

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/64032>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

**Recursive partitioning based approaches for
low back pain subgroup identification in
individual patient data meta-analyses**

by

Dipesh Mistry

A thesis submitted in partial fulfilment of the requirements for the
degree of
Doctor of Philosophy in Health Sciences

The University of Warwick, Warwick Medical School
August 2014

Contents

List of Tables	vi
List of Figures	ix
Acknowledgements	xi
Declaration	xiii
Abstract.....	xiv
List of Abbreviations.....	xv
Part I - Context and Current Practice	1
1 Brief Introduction.....	2
1.1 Introduction	2
1.2 Project Background.....	3
1.3 Research Objectives	3
1.4 Thesis Structure.....	5
2 Background	8
2.1 Introduction	8
2.2 Definition.....	8
2.3 Epidemiology	10

2.4 Costs	10
2.5 Available Treatments	11
2.6 Outcome Measures	14
2.7 Effect sizes in positive studies	19
2.8 Discussion	21
 3 Subgroup Analyses in Clinical Trials	 23
3.1 Introduction	23
3.2 Subgroup analysis concept.....	25
3.3 What is a subgroup analysis?	26
3.4 Statistical concepts and issues with subgroup analyses.....	27
3.5 Subgroup analyses guidance	36
3.6 Issues with current recommendations	39
3.7 Discussion	41
 4 Systematic review of subgroup analyses in non-specific low back pain trials	 43
4.1 Introduction	43
4.2 Methods.....	44
4.3 Results	48
4.4 Discussion	52
 Part II - Statistical Methodology	 66
5 Review of subgroup methodology.....	67
5.1 Introduction	67
5.2 Methods to identify subgroups with high or low outcome.....	68
5.3 Methods to identify subgroups with high or low treatment effects.....	72
5.3.1 Single factors	72

5.3.2 Multiple factors.....	75
5.4 Subgroup analysis methods for IPD from trials	78
5.5 Discussion	79
6 Introduction to Recursive Partitioning	82
6.1 Introduction	82
6.2 Tree Based Methods.....	83
6.2.1 Growing an Initial Fully Grown Tree	85
6.2.2 Tree growing example	86
6.2.3 Node splitting.....	87
6.2.4 Pruning.....	90
6.2.5 Selecting the optimal sub-tree	93
6.2.6 Interpretation of the optimal tree T^*	96
6.3 Advancements of recursive partitioning methodology to detect moderators of treatment effect.....	96
6.3.1 IT method.....	98
6.3.2 STIMA method	101
6.3.3 SIDES method.....	106
6.4 Discussion	113
7 Simulation study to evaluate tree based methods.....	116
7.1 Introduction	116
7.2 Simulation study setup	117
7.3 Final trees grown by methods	123
7.4 Simulation Results.....	128
7.5 Discussion	131

8	Extension of recursive partitioning approaches to IPD meta-analysis	137
8.1	Introduction	137
8.2	Introduction to IPD meta-analyses	139
8.3	Statistical methods for IPD subgroup meta-analyses	140
8.4	Proposed extension of tree methods	144
8.5	Simulation study setup	150
8.6	Simulation study results.....	151
8.7	Discussion	154
9	Further development of the IPD-SIDES method.....	168
9.1	Introduction	168
9.2	Investigating the drop in performance of the IPD-SIDES method	169
9.3	Simulation Study	173
9.4	Simulation Study 1 results	174
9.5	Simulation Study 2 results	174
9.6	Discussion	175
	Part III - Application and Summary.....	181
10	Application of IPD-IT and IPD-SIDES to real data	182
10.1	Introduction.....	182
10.2	Description of the pooled dataset.....	183
10.3	Methods.....	184
10.4	Results.....	187
10.5	Discussion.....	193
11	Discussion, Further Work & Conclusions	199
11.1	Introduction.....	199

11.2 Discussion and Further Work	199
11.3 General recommendations	205
11.4 Conclusions	207
11.4.1 Addressing the research objectives of this thesis.....	207
11.4.2 Contributions to the literature	208
Bibliography	210
 Part IV - Appendices.....	 224
Appendix A: Systematic review paper published in the Spine Journal.....	225
Appendix B: Systematic review search terms	238
Appendix C: Matrix form of the general expressions required to generate data for the simulation study	239

List of Tables

2.1	Comparison of interventions from high-quality RCTs at short term (3 month) and long term (12 month) follow-up	18
3.1	Summary table of treatment group by employment status interaction	28
4.1	Inclusion and exclusion criteria used to select papers	45
4.2	Summary of included papers ordered by subgroup quality assessment	56
4.3	Summary of excluded papers	63
7.1	2 x 2 x2 table of within cell means for T , X_1 and X_2	121
7.2	Simulation results for the IT method. Results display % of correctly identified final trees	132
7.3	Simulation results for the STIMA method. Results display % of correctly identified final trees	133
7.4	Simulation results for the SIDES method. Results display % of correctly identified candidate subgroups	134
7.5	Simulation results presenting the proportion (%) of the different tree sizes obtained for a range of sample sizes to investigate STIMA method when a very large and a large interaction effect are present	135
8.1	Simulation results for the IPD-IT method for Model A (Null model) when there is small (0.1) between-study variation. Results display % of correctly identified final trees	156

8.2	Simulation results for the IPD-IT method for Model B (Fixed-effect model) when there is small (0.1) between-study variation. Results display % of correctly identified final trees	157
8.3	Simulation results for the IPD-IT method for Model C (Mixed-effect model) when there is small (0.1) between-study variation. Results display % of correctly identified final trees	158
8.4	Simulation results for the IPD-IT method for Model A (Null model) when there is large (0.9) between-study variation. Results display % of correctly identified final trees	159
8.5	Simulation results for the IPD-IT method for Model B (Fixed-effect model) when there is large (0.9) between-study variation. Results display % of correctly identified final trees	160
8.6	Simulation results for the IPD-IT method for Model C (Mixed-effect model) when there is large (0.9) between-study variation. Results display % of correctly identified final trees	161
8.7	Simulation results for the IPD-SIDES method for Model A (Null model) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups	162
8.8	Simulation results for the IPD-SIDES method for Model B (Fixed-effect model) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups	163
8.9	Simulation results for the IPD-SIDES method for Model C (Mixed-effect model) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups	164
8.10	Simulation results for the IPD-SIDES method for Model A (Null model) when there is large (0.9) between-study variation. Results display % of correctly identified final subgroups	165
8.11	Simulation results for the IPD-SIDES method for Model B (Fixed-effect model) when there is large (0.9) between-study variation. Results display % of correctly identified final subgroups	166
8.12	Simulation results for the IPD-SIDES method for Model C (Mixed-effect model) when there is large (0.9) between-study variation. Results display % of correctly identified final subgroups	167
9.1	Simulation output for when there is a $t \times v1$ interaction effect of 1.5 and total sample size $N=5000$	177
9.2	Simulation results for the IPD-SIDES method with a restriction on the differential effect p-value ($p \leq 0.10$) when there is small (0.1) between-	177

	study variation. Results display % of correctly identified final subgroups	
9.3	Simulation results for the IPD-SIDES method with a restriction on the differential effect p-value ($p \leq 0.10$) when there is small (0.9) between-study variation. Results display % of correctly identified final subgroups	178
9.4	Simulation results for the modified IPD-SIDES with a restriction on the differential effect p-value ($p \leq 0.10$) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups	179
9.5	Simulation results for the modified IPD-SIDES with a restriction on the differential effect p-value ($p \leq 0.10$) when there is small (0.9) between-study variation. Results display % of correctly identified final subgroups	180
10.1	Cochrane collaboration Risk of Bias assessment of included trials	188
10.2	Summary of demographic and baseline data from pooled dataset	189
10.3	Summary of short-term outcome data and change from baseline to short-term follow-up outcome data	190
10.4	Subgroups identified by the unmodified IPD-SIDES method when applied to the change from baseline to short-term MCS outcome data	196
10.5	Subgroups identified by the unmodified IPD-SIDES method when applied to the change from baseline to short-term PCS outcome data	196
10.6	Subgroups identified by the modified IPD-SIDES method when applied to the change from baseline to short-term MCS outcome data	198
10.7	Subgroups identified by the modified IPD-SIDES method when applied to the change from baseline to short-term PCS outcome data	198

List of Figures

3.1	Quantitative Interaction (treatment effect in the same direction)	31
3.2	Qualitative Interaction (treatment effect in different directions)	31
3.3	Probability of falsely detecting at least one significant interaction at the 5% significance level for independent hypothesis tests	33
4.1	Flow diagram	49
6.1	Example of a regression tree	87
6.2	Example of STIMA tree for one-way interaction	104
6.3	Example of the SIDES procedure with two levels	115
7.1	Final tree produced by the IT method for a single one-way interaction	124
7.2	Final tree produced by the IT method for two single one-way interactions	124
7.3	Final tree produced by the STIMA method for a single one-way interaction	125
7.4	Final tree produced by the STIMA method for two single one-way interactions	126
7.5	Final tree produced by the SIDES method for a single one-way interaction	126
7.6	Final tree (A) produced by the SIDES method for two single one-way interactions; illustrated using a single tree	127
7.7	Final tree (A) produced by the SIDES method for two single one-way interactions; illustrated using a two separate tree	127

7.8	Final tree (B) produced by the SIDES method for two single one-way interactions; illustrated using a single tree	128
7.9	Final tree (B) produced by the SIDES method for two single one-way interactions; illustrated using two separate trees	128

Acknowledgements

I dedicate this thesis to Lord Swaminarayan and my Guru, His Divine Holiness Pramukh Swami Maharaj. Everything I have achieved in life and the completion of the work in this thesis would not be possible without their divine blessings and inspiration.

I am forever indebted to my loving parents, Balwantlal Mistry and Sharda Mistry, for fully supporting me in every way possible and for believing in me in everything that I have ever done. Words cannot describe my gratitude and heartfelt appreciation to my loving wife Jayshree, my siblings (Heena, Kavita and Pankaj) and friends for all their support, patience and words of encouragement throughout this journey.

I would like to take this opportunity to express my deepest appreciation and sincere gratitude to both of my supervisors, Professor Martin Underwood and Professor Nigel Stallard for their constant guidance, encouragement, motivation and support throughout my PhD. I am extremely grateful for the invaluable knowledge I have gained from them and I cannot thank them enough for the time and patience through it all. It has been a great privilege and absolute pleasure working with them and under their supervision.

Finally, a massive thank you goes to my office colleagues for always being there to advise me and humour me to get me through each day and make it a very fun three years.

Declaration

I declare that this thesis is my own work, except where I have stated otherwise, and I can also confirm that this thesis has not been submitted elsewhere for a degree at another university. The systematic review presented in Chapter 4 has been published in the Spine Journal (published January 2014). The work contained in the systematic review has been primarily composed by myself except for the search strategy (Appendix A) and list of potential papers which was kindly provided by Dr Gurung who performed the systematic review for the main funded project.

Abstract

This thesis presents two novel approaches for performing subgroup analyses or identifying subgroups in an individual patient data (IPD) meta-analyses setting. The work contained in this thesis originated from an important research priority in the area of low back pain (LBP); identifying subgroups that most (or least) benefit from treatment. Typically, a subgroup is evaluated by applying a statistical test for interaction between a baseline characteristic and treatment. A systematic review found that subgroup analyses in the area of LBP are severely underpowered and are of a rather poor quality (Chapter 4). IPD meta-analyses provide an ideal framework with improved statistical power to investigate and identify subgroups. However, conventional approaches to subgroup analyses applied in both a single trial setting and an IPD setting have a number of issues, one of them being that subgroups are typically investigated one at a time. As individuals have multiple characteristics that may be related to response to treatment, alternative statistical methods are required to overcome the associated issues. Tree based methods are a promising alternative that systematically search the entire covariate space to identify subgroups defined by multiple characteristics. In this work, a number of relevant tree methods, namely the Interaction Tree (IT), Simultaneous Threshold Interaction Modelling Algorithm (STIMA) and Subpopulation Identification based on a Differential Effect Search (SIDES), were identified and evaluated in a single trial setting in a simulation study. The most promising methods (IT and SIDES) were extended for application in an IPD meta-analyses setting by incorporating fixed-effect and mixed-effect models to account for the within trial clustering in the hierarchical data structure, and again assessed in a simulation study. Thus, this work proposes two statistical approaches to subgroup analyses or subgroup identification in an IPD meta-analysis framework. Though the application is based in a LBP setting, the extensions are applicable in any research discipline where subgroup analyses in an IPD meta-analysis setting is of interest.

List of Abbreviations

AID	Automatic interaction detection
CART	Classification and regression trees
GAM	Generalized additive models
GLM	Generalized linear model
GPE	Global perceived effect
GUIDE	Generalized unbiased interaction detection and estimation
IDET	Intradiscal electrothermal therapy
IPD	Individual patient data
IPD-IT	Individual patient data interaction tree
IPD-SIDES	Individual patient data subgroup identification based on a differential effect search
IT	Interaction tree
LBP	Low back pain
MCS	Mental component score
MFPI	Multivariable fractional polynomials interaction
MIC	Minimum important change
ML	Maximum likelihood
MVK	Modified Von Korff
NHS	National health service
NICE	National institute for health and care excellence

NIHR	National institute of health research
NRS	Numerical rating scale
NSAID	Non-steroidal anti-inflammatory drugs
ODI	Oswestry disability index
PCS	Physical component score
PIRFT	Percutaneous Intradiscal radiofrequency thermocoagulation
PSE	Pain self-efficacy
RCT	Randomised controlled trial
REML	Restricted maximum likelihood
RMDQ	Roland Morris disability questionnaire
RTA	Regression trunk approach
SD	Standard deviation
SE	Standard error
SIDES	Subgroup identification based on a differential effect search
SMD	Standardised mean difference
SSE	Sum of squared errors
SSRI	Selective serotonin reuptake inhibitors
STEPP	Subpopulation treatment effect pattern plot
STIMA	Simultaneous threshold interaction modelling algorithm
SVM	Support vector machines
TENS	Transcutaneous electrical nerve stimulation
UK	United Kingdom
USA	United States of America
VAS	Visual analogue scale
VKS	Von Korff scale

PART I

Context and Current Practice

Chapter 1

Brief Introduction

1.1 Introduction

The main purpose of this thesis is to develop and evaluate an innovative approach for identifying subgroups in individual patient data (IPD) meta-analyses within the area of low back pain (LBP). Although this is the ultimate goal, there are several important steps that need to be taken along the way to help achieve this goal. These steps include a systematic review of the quality of subgroup analyses in LBP randomized controlled trials (RCTs), a review of the methods used to perform subgroup analyses in general and the comparison of identified relevant subgroup analysis methods in a single trial setting. Evaluating candidate methods in a single trial setting will help select the best method(s) to take forward and extend to an IPD subgroup meta-analyses setting.

At the outset, it is important to put this PhD project into context and make it clear as to where this work originated from and why it was funded. Therefore, this chapter will start by briefly detailing the background for the overall project and origins of the PhD funding. The primary and secondary research objectives of this thesis will then be presented. Finally, an overview of the thesis structure will be given.

1.2 Project Background

The funding for this PhD comes from a National Institute of Health Research (NIHR) programme grant applied for by both the supervisors of this PhD; Professor Martin Underwood (Principal Investigator) and Professor Nigel Stallard (Co-applicant). The funding was for a programme grant on the identification of subgroups of patients that most benefit from therapist delivered interventions for the management of LBP. This project involved collecting a repository of individual patient data from the several existing RCTs that have tested similar therapist delivered interventions, and then performing subgroup analyses of the pooled trial dataset. The outcome of the subgroup analyses would determine whether or not patients presenting with non-specific low back pain can be matched to a treatment that is best suited to them. Any identified subgroups that most benefit will have increased effectiveness compared to the average.

The programme grant also funded this PhD to look into the development of subgroup analysis methodology specifically in the area of low back pain research. The work, therefore, sits clearly within the field of back pain research, but with a strong focus on methodological development in statistical analysis.

1.3 Research Objectives

The work in this thesis is motivated by research priorities in the LBP community; in particular, identifying subgroups of patients who most benefit from therapist delivered interventions. Typically single trials are designed such that the sample size will provide sufficient power for a main effect to be detected in the primary outcome only.

Therefore splitting the total sample into smaller subgroups will severely under-power any secondary subgroup analyses for detecting true subgroup effects i.e. more likely to yield false negative findings. Despite lacking power, subgroup analyses can still be

performed as a purely exploratory piece of work to generate hypotheses to be tested in future research.

There are plenty of LBP data from positive studies that test similar interventions and use similar outcome measures. Thus a cost-effective way of performing subgroup analyses that substantially improves the issue of power is by pooling the IPD from these trials and performing subgroup meta-analyses. Ordinary meta-analyses that synthesize aggregated data from several similar studies are a popular form of analysis. It has been used for many years; hence its methodology is very well established. IPD meta-analyses on the other hand use the original individual patient data from each of the studies, which makes the analyses rather more complex. However, of the two approaches, the advantage of IPD meta-analyses is that patient level covariates can be fully investigated which is therefore ideal for performing subgroup analyses. Performing IPD meta-analyses has only recently become popular and therefore its methods are not as well established. Typically, subgroup analyses in both single trials and in IPD meta-analyses use interaction tests and only test for one interaction at a time; they do not consider the multiple characteristics of patients. Thus, there is a need to develop IPD subgroup meta-analyses methodology that incorporates the multiple characteristics of patients when defining and identifying subgroups. This highlights the primary research objective of the work in this thesis.

Primary research objective: The primary research objective of this thesis is to develop and evaluate an innovative approach or approaches for identifying subgroups in IPD meta-analyses within the area of low back pain.

Secondary research objectives: In order to attain the primary research objective, a number of important steps in the form of secondary research questions were

addressed. These steps are primarily to do with evaluating both the clinical and statistical aspects of subgroup analyses in a single trial setting. By doing so, this helps develop a firm understanding of subgroup analyses as a whole before moving forward to think about methodological extensions to an IPD subgroup meta-analyses setting. To be more precise, the secondary research questions were:

- 1) What are the current recommendations for performing subgroup analyses in RCTs?
- 2) What is the quality, conduct and reporting like of secondary subgroup analyses performed of RCTs of therapist delivered interventions for the management of non-specific LBP?
- 3) What alternative methods can be used to perform subgroup analyses in RCTs?

These secondary research questions helped to systematically work towards meeting the primary objective. These questions were answered in the above order and thus formed the structure of this thesis.

1.4 Thesis Structure

Due to the multidisciplinary nature of this project, the content of this thesis has been separated into three parts with the intention of maintaining clarity and focus. More specifically, this was done to make the statistical and clinical aspects more distinct, at the same time allowing the statistical component to be presented in a more general setting. The three parts of this thesis are as follows:

Part 1: Context and Current Practice

Chapter 1 - Brief introduction

Chapter 2 – Background

Chapter 3 – Review of current subgroup analysis guidance for trials

Chapter 4 – Systematic review of subgroup analyses performed in LBP RCTs

Part 2: Statistical Methodology

Chapter 5 – Review of generic subgroup methodology

Chapter 6 – Introduction to recursive partitioning methodology

Chapter 7 – Simulation study comparing existing tree based methods based on recursive partitioning methodology

Chapter 8 - Development of tree based method using recursive partitioning methodology to identify subgroups in individual patient data meta-analyses

Chapter 9 – Further development of tree based methods

Part 3: Application and Summary

Chapter 10 – Application of proposed method to real data

Chapter 11 – Discussion, further work and conclusions

Chapters in Part I of the thesis will focus more on the clinical aspect of the project. At first, it is important to develop a firm understanding about what LBP is, what the current recommendations are for the management of LBP, how LBP is currently measured and why the identification of subgroups is a high research priority in the area of LBP. Thus, a detailed background of the clinical problem will be given in Chapter 2. Although several secondary subgroup analyses have been performed in the LBP literature, it is important to review the quality of these reported analyses. However, prior to doing so, it is necessary to understand the statistical challenges faced when performing subgroup analyses e.g. lack of power, multiplicity issue, and learn what recommendations have been made to ensure that the reported results are credible. For this reason, a review of the statistical challenges and current recommendations for performing subgroup analyses in RCTs in general is presented in Chapter 3. Thereafter, the final chapter in Part I will present a systematic review of

subgroup analyses performed in RCTs of therapist delivered interventions for the management of non-specific LBP. The content of this systematic review has been published in the Spine Journal (1).

Chapters in Part II of the thesis focus on the statistical aspects of the project. This part of the thesis will initially review subgroup analysis methodology in general in Chapter 5, where a host of methods that can be used to perform subgroup analyses will be described. Chapter 5 identifies tree based approaches that are based on recursive partitioning methodology as a promising approach to subgroup analyses in the context of RCTs. Chapter 6 will therefore go on to introduce the recursive partitioning methodology along with a description of several advanced variants that use this methodology to identify subgroups that moderate treatment effect. A simulation study will then be performed in Chapter 7 to compare the tree based methods described in Chapter 6 in a single trial setting. The candidate methods identified from Chapter 7 will then be developed for application to individual patient data. Therefore, Chapter 8 will detail the development of the proposed statistical method for identifying subgroups in an IPD subgroup meta-analysis including a simulation study. Finally in Part II of the thesis, further development of the proposed extensions evaluated in Chapter 8 will be described and assessed.

Finally, in Part III I will focus on applying the proposed methodology to real data collected for the main funded project. The application of the proposed methods will be presented in Chapter 10. A detailed discussion, recommendations for further work and the final conclusions will be presented thereafter in Chapter 11.

Chapter 2

Background

2.1 Introduction

Low back pain (LBP) is a musculoskeletal disorder that is extremely common and experienced by most adults at some point in their life (2, 3). It is common for a person with LBP to experience recurring episodes over time which in turn will affect their quality of life as well as their well-being (4, 5). Consequently, it has a big health and social impact, causing an enormous economic burden on those in search of treatment, their families and to society as a whole (6-8). The effective management of LBP is therefore a major concern. This chapter will initially define non-specific LBP along with its epidemiology and costs. The available recommended treatments and the most commonly used outcome measures of LBP will then be described. Finally, the effect sizes in positive studies will be discussed highlighting subgroup analyses as an important research priority.

2.2 Definition

In general, LBP is defined clinically as the occurrence of pain and discomfort located in the lower region of the back (also referred to as the lumbosacral area), that is, the area between the bottom of the rib cage and the buttock creases (9). In some instances, pain

in the upper legs may also be experienced in addition to the LBP. Here the term 'pain' is commonly defined as being an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage (10). The term 'non-specific' LBP simply refers to when pain is experienced in the lower back region but it is not possible to pin-point the exact cause of the LBP or attribute it to a known pathology (e.g. fracture, infection, malignancy specific cause etc). The set of symptoms for non-specific LBP include tension, soreness and/or stiffness in the lower back region attributed to a combination of several structures in the back, including the discs, joints and connective tissues (9). Even though non-specific LBP is typically defined in this way, there is a clear misconception as to what it actually means. It is defined here as being a disease or a condition when in actual fact it is a set of symptoms.

The duration of an episode of non-specific LBP can be classified as being acute, sub-acute or chronic. Conventionally, acute LBP is when the duration of an episode continues for no more than six weeks, sub-acute LBP is when the duration of an episode lasts between six and twelve weeks and finally chronic LBP is when the duration of an episode lasts for more than 12 weeks (11). Although classifying patients in this manner may help with the diagnosis and management of LBP, it is possible that the approach in itself may not help with treatment decisions. For example, it might just be that the duration of symptoms is not a strong predictor of outcome and does not predict response to treatment at all (12). Also the symptoms of LBP may vary over time therefore making it difficult to classify the patient as being either an acute case or a chronic case of LBP. Applying such artificial constructs to patient symptoms may just be an oversimplification of what is happening in reality. Instead, it may be more useful to investigate diagnostic classification criteria whereby the large population of non-specific LBP patients are further separated into smaller more homogeneous subgroups

based on some valid criteria e.g. history, muscular stability and flexibility. Such an approach may allow patients to be treated more effectively.

2.3 Epidemiology

Non-specific LBP is a common problem that can be quite disabling and is experienced by a majority of the UK adult population at some point in their life. A recent systematic analysis of the global burden of disease highlighted LBP as the leading cause contributing the most to the overall years lived with disability (13). It is difficult to estimate the prevalence of LBP as it tends to vary from study to study. However a systematic review of studies for LBP prevalence from 1966 to 1998 reported that there was 12% - 33% point prevalence, 22% - 65% 1-year prevalence and up to 84% life time prevalence (14). LBP is a recurrent problem that fluctuates over time. The majority of LBP patients, around 70%, will experience at least one recurrent episode within a 12 month period (15, 16). Estimates suggest that every year, around a third of the adults in the UK experience acute LBP but the symptoms often resolve quickly. A large number of patients do not consult the NHS and instead opt to self-manage their LBP. It is estimated that up to 20% approach their general practitioner (GP) about their problem (17). This works out to be around 2.6 million additional consultations every year (18), of which nearly 75% still have symptoms after a year and around 30% develop prolonged recurrent LBP (19, 20).

2.4 Costs

Health care costs are referred to as being direct financial costs whereas production loss costs and the cost of insurance as a result of injury are referred to as indirect costs. In all developed countries, the direct and indirect financial costs of LBP are considerably

large. However, it is not possible to make any direct comparisons of the cost of LBP due to varying health and social systems (21).

Based on the most recent UK cost of illness study, the direct healthcare costs to the National Health Service (NHS) due to LBP were in the region of £1,632M (3). Approximately a third of this (£565M) was from non-NHS based health care costs which were primarily due to the use of private sector services (physiotherapists, osteopath's, chiropractors, acupuncturists and others). The indirect cost estimates on the other hand vary due to the type of model used but could have been as much as £10,668M (3). It is important to remember here that these estimates are based on estimates made in 1998. Since then, there has been inflation as well as a dramatic change in the economy and therefore it is likely that the current direct and indirect financial costs for LBP are higher than those figures published in 1998.

2.5 Available Treatments

There are a wide variety of available treatments for the management of persistent non-specific LBP that have been recommended in guidelines in the UK (National Institute for Health and Care Excellence (NICE) guidelines), Europe (European Guidelines-COST B13) and America (American College of Physicians and American Pain Society guidelines) (9, 22, 23). When there has been evidence to suggest that an intervention is beneficial, it is often not possible to distinguish whether the benefit is because of the intervention itself (specific effect) or as a result of the therapist's delivery of the intervention (non-specific effect). For this reason, each of the guidelines clearly clarify that any guidance for the management of LBP is based on the assumption that the effects are purely down to the intervention package alone.

All patients presenting with non-specific LBP should first be given advice and information to encourage self-management of their problem (9). This involves informing the patient of the characteristics and symptoms of non-specific LBP along with advice to remain physically active and to persist in their normal daily activities as much as they can. This is standard practice for all patients. Next, all treatment options should be clearly communicated by the healthcare professional to the patient. Subsequently, a decision is made jointly by the patient and the healthcare professional taking into account the patient's treatment preference and needs when selecting one of the following recommended treatments:

- Physical activity
- Exercise
- Manual therapy
- Combined physical and psychological treatment

From the above, the combined physical and psychological treatment option is considered a bit later on in the care pathway after the other three simpler treatments and before any suggestions of surgery. If the chosen treatment package is not successful then another package should be offered. A person should only be referred for surgery if they have completed the best possible package that is suited to them and their severe LBP still remains. In such an instance, the patient should be referred to a service that specializes in spinal surgery.

Drug Treatments

In the first instance, paracetamol should be offered as initial medication. If paracetamol fails to provide effective relief then weak opioids, non-steroidal anti-inflammatory drugs (NSAIDs) or both can be offered. It is essential that the potential side-effects,

particularly in the elderly population, based on each individual are taken into account. This should then be explained to the patient so that the drug can be prescribed based on patient preference.

Treatments not recommended

There are a number of treatments for which there is not enough adequate evidence to know whether or not they work for the management of non-specific LBP and are therefore not recommended. They include:

- Laser therapy
- Therapeutic ultrasound
- Interferential therapy
- Transcutaneous electrical nerve stimulation (TENS)
- Lumbar supports
- Spinal traction
- Therapeutic matter injected into the back
- Selective serotonin reuptake inhibitors (SSRI)
- Intradiscal electrothermal therapy (IDET)
- Percutaneous intradiscal radiofrequency thermocoagulation (PIRFT)
- Radiofrequency facet joint denervation

The recommended treatments and best practice advice described above are primarily from the recommendations of the more recent UK (2009) and American (2007) guidelines which are based on more current evidence. The UK and American guidelines make rather similar recommendations for the management of non-specific LBP but have a few subtle differences when compared to the European guidelines (9, 22, 23). For example, the European guidelines do not recommend acupuncture and massage

therapy for the management of chronic LBP; whereas NICE and American guidelines do. However, these differences are most likely due to the lack of evidence available in the literature at the time when forming the European (2004) guidelines.

2.6 Outcome Measures

Many of the earlier trials investigating interventions for non-specific LBP used different outcome measures which made between study comparisons rather problematic. It was then decided by a group of international investigators that a main set of 5 domains should be used in all trials to standardize the outcome measures and to better facilitate the comparability amongst studies (24). The five domains are (i) pain symptoms, (ii) back related function, (iii) Generic well-being, (iv) Disability and (v) Satisfaction with care. A brief description of the most common outcomes used within each of these domains is given below:

(i) Pain Symptoms

- *Troublesomeness/Bothersomeness* - this is typically a single question asking the patient how troublesome or bothersome their LBP is. Patients can answer the question by selecting one of the five possible response categories which are Not at all, Slightly, Moderately, Very much and Extremely (25, 26).
- *Pain Severity* – the severity or intensity of back pain is often measured using either a numerical rating scale (NRS) on a scale of 0 (No pain) to 10 (Worst possible pain) or a visual analogue scale (VAS) on a scale of 0 (No pain) to 100 (Worst imaginable pain) (27). The NRS is a one-dimensional measure that can be administered either graphically or verbally. The patient is required to indicate a whole number on an 11-point scale from 0 to 10 that best reflects their level of pain. The number indicated by the patient is then recorded by the person administering the questionnaire. The VAS on the other hand is a

continuous one-dimensional scale that can take the form of either a horizontal or vertical line that is 100mm (10cm) in length. The VAS is administered to the patient for self-completion where the patient is required to draw a line perpendicular to the VAS line that best reflects their level of pain.

- *Frequency* – this is a single question that enquires about how frequently the patient experiences LBP. The most commonly used question requires that the patient selects the category that best reflects their frequency of pain. The number of available response categories varies but the two extremes are usually ‘Never’ and ‘Always’. An alternative measure of frequency might be the number of days of pain over a certain duration e.g. over the past 4 weeks.

(ii) Back related function

- *Roland Morris Disability Questionnaire (RMDQ)* – the RMDQ questionnaire consists of twenty-four items and measures back related disability (28). The patient is required to tick any of the twenty-four items that applies to them. The final score is computed by simply counting the number of ticked boxes. The final score is on a scale of 0-24 where a lower score indicates less severe disability. However a problem with this questionnaire is that it has not got any ‘No’ tick boxes. This makes it impossible to distinguish whether an un-ticked box is genuinely un-ticked or has been missed by the patient.
- *Oswestry Disability Index (ODI)* – the ODI questionnaire consists of ten questions related to everyday activities of daily living, where each question is on a scale of 0 (no disability) to 5 (worst disability) (29). The scores for each of the questions are then summed and converted to a scale of 0-100% where a lower score indicates less disability.
- *Von Korff Scale (VKS)* - the modified Von Korff questionnaire measures both pain and disability using six questions, with each question having a scale of 0

(no pain/disability) to 10 (worst pain/disability). The first three questions are used to compute a measure of disability whereas the last three are used to measure pain. The scores are then transformed to a 0-100% scale where a lower score represents less pain or disability (30).

(iii) Generic well-being

- *SF-12 or SF-36* - the SF-12 or SF-36 are commonly used as a measure of health-related quality of life. The SF-12 questionnaire consists of 7 questions with 12 items in total whereas the SF-36 questionnaire consists of 11 questions with 36 items in total (31, 32). An algorithm is then applied to the item responses to create eight subscales that measure functional health and well-being. These eight subscales measure physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional and mental health. These eight subscales are then used to create two aggregated summary measures of physical and mental health. These are measured on a scale of 0-100 where a lower score indicates poorer physical or mental functioning.
- *EQ-5D* – the EQ-5D is an outcome that measures the health-related quality of life of a patient. The questionnaire consists of two parts. The first part asks questions based on five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression), hence why it is also referred to as the EQ-5D, where each dimension consists of three possible options (33). An algorithm is then applied to the responses to compute a healthy utility score that is usually on a scale of 0 (death) to 1 (perfect health state). It is also possible to obtain a negative utility score which suggests that a patient's quality of life is worse than death. The second part of the questionnaire consists of a visual analogue scale (VAS) that has a scale of 0 (worst imaginable health state)

to 100 (best imaginable health state). The patient is required to mark on the scale how good or bad their health state is today.

(iv) Disability

- *Days off work* – this question simply asks the patient for the number of days off from work due to LBP over a certain period of time. The period of time for which this information is required varies from study to study e.g. number of days off from work during the past 4 weeks? Or number of days off from work during the past 3 months?

(v) Satisfaction

- *Participant satisfaction* – this is a single item question, as recommended by the international low back forum, that enquires about how satisfied the patient is with the effectiveness of the treatment they have received (24). There may be variations of the response categories available for this outcome measure. Typically the categorical responses are on a five point Likert scale ranging between the two extremes of 'Very dissatisfied' and 'Very satisfied'.
- *Specific health transition question* – this single question asks the patient how they feel since they were last assessed. The patient can choose one of several categorical responses on a Likert scale that best reflects how they feel. Responses include 'are you much worse', 'a bit worse', 'the same', 'a bit better' or 'a lot better'.

Table 2.1 – Comparison of interventions from high-quality RCTs at short term (3 month) and long term (12 month) follow-up

Study	Control	Intervention	Mean Difference in RMDQ (95% CI); SMD	
			3 month	12 month
UK BEAM (2004)	GP care	Exercise	1.36 (0.63, 2.10); 0.34	0.39 (-0.41, 1.19); 0.10
		Manipulation	1.57 (0.82, 2.32); 0.39	1.01 (0.22, 1.81); 0.25
		Manipulation plus exercise	1.87 (1.15, 2.60); 0.47	1.30 (0.54, 2.07); 0.33
ATEAM (2008)	Usual care	Massage	1.96 (0.74, 3.18); 0.39	0.58 (0.77, 1.94); 0.12
		Alexander technique (6 sessions)	1.71 (0.47, 2.95); 0.34	1.40 (0.03, 2.77); 0.28
		Alexander technique (12 sessions)	2.91 (1.66, 4.16); 0.58	3.40 (2.03, 4.76); 0.68
BeST (2010)	Advice only	Cognitive behavioural therapy	1.10 (0.38, 1.71); 0.22	1.30 (0.56, 2.06); 0.27
York Yoga (2011)	Usual care	Yoga	2.17 (1.03, 3.31); 0.50	1.57 (0.42, 2.71); 0.36

2.7 Effect sizes in positive studies

Due to the severe debilitating effect of LBP and the enormous economic burden induced, substantial amounts of time and money have been expended researching various interventions since the mid-1990s. There are now several randomized controlled trials (RCTs) of different interventions that have proven to be effective with some of these also being cost-effective. To give some examples, Table 2.1 details the results of some high-quality RCTs, namely the UK BEAM trial, ATEAM, BeST and York Yoga trial (34-37). The magnitude of the effect sizes presented in Table 2.1 may be important at the population level however such effect sizes are much smaller than the general consensus of the minimum important change (MIC) at the individual level (38).

Not all studies use the same outcome measure and so it is rather difficult to compare effect sizes amongst studies. To overcome this, effect sizes are typically standardized by dividing it by the standard deviation of the outcome measure at baseline to convert it to a standard unit of measure. The standardized measure thus reflects how many standard deviations the outcome measure changes from baseline up to the follow-up time point of interest. Standardized effect sizes, also referred to as standardized mean differences (SMD), are commonly described as being small (0.2), moderate (0.5) and large (0.8), as described by Cohen (39). In general, the effect sizes for the different available interventions in the area of LBP are of a similar magnitude; demonstrating, at best, small to moderate standardized mean effect sizes. Nevertheless, there is no clear indication of a substantial reduction in the health and social impact of LBP such as time lost from work and the burden of disability. This is clearly evident in a recent study of the UK global burden of disease which suggests that there was a 12% increase in the years lived with disability in the UK between 1990 and 2010 (40). The question then remains as to whether or not these treatments are really worthwhile.

There are plenty of useful LBP data from different trials to show these approaches work relatively well for the general LBP population as a whole. Thus, one approach to maximize treatment benefit for the individual would be to utilize these data to identify those who most benefit from different treatments. Identifying such subgroups would help enable clinicians to target treatment accordingly which in turn would improve individual patient care. For this reason, discovering which patients most benefit from treatment is highlighted as a key recommendation by the UK National Institute for Health and Care Excellence (NICE) LBP guidelines (9).

Typically, LBP RCTs are designed and powered to test a main hypothesis in terms of a primary outcome measure. Thus, any secondary subgroup analyses will be substantially underpowered to detect the same effect size as that of the primary analysis. Notwithstanding this lack of statistical power, secondary subgroup analyses still have the potential to identify important patient subgroups, provided that the quality of subgroup analyses is of a high standard. Any identified subgroups would need to be further investigated in future trials. Numerous papers in the LBP literature claim to have performed subgroup analyses however the overall quality of these analyses is unknown. If the quality, conduct and reporting of subgroup analyses is poor, then it will have false implications on future research and future treatment. For example, assuming a paper reports a positive finding for a subgroup analysis that has been incorrectly conducted; another investigator may thereafter naively use the reported finding to conduct a future piece of research. Not only would this be a waste of the investigators time and money, but this would also increase the chances of more poor quality findings being reported in the literature. Thus it is very important to ensure that the quality of subgroup analyses in the area of LBP is of a good standard by following available proposed subgroup analyses guidelines (41-44).

2.8 Discussion

This chapter gave an insight to the background of non-specific LBP as well as briefly describing the recommended available treatments based on the evidence to date. In general, the recommended treatment options available all show small to moderate effect sizes in improving non-specific LBP and there is no indication that a particular treatment is superior compared to the others.

A number of Cochrane reviews in the area of non-specific LBP have been performed either comparing various treatments with each other or compared to usual care (45-49). Typically, these reviews provide a clear and consistent message that the available treatments are only able to demonstrate small to moderate positive change when compared to usual care but there is no indication of a single specific treatment being superior. Thus the evidence suggests that the available treatments are relatively effective in the general LBP population as a whole. However, it may well be that certain subgroups of individuals benefit more (or less) from these treatments. As we are striving towards more individualized patient care, it is important that secondary analyses be performed to identify these subgroups and extract as much information as possible. This was one of the key recommendations made in the UK, European and American guidelines for future research for the management of non-specific LBP (9, 22, 23).

Over the years, there have been several studies evaluating different interventions for the management of LBP that have also performed secondary subgroup analyses. Acknowledging the fact that subgroup analyses are severely underpowered, it is not entirely clear what the quality of subgroup analyses is like in the LBP literature as a whole. It is therefore of interest to perform a systematic review of the quality, conduct and reporting of these secondary subgroup analyses. As most of the positive evidence

is for therapist delivered interventions, the systematic review as well as the rest of this PhD will be based on therapist delivered interventions and finding subgroups in this area.

Prior to reviewing subgroup analyses in the LBP literature, it is important to review the statistical aspects and current proposed guidelines or recommendations for performing subgroup analyses in RCTs in general. Therefore, the next chapter will review current proposed recommendations for performing subgroup analyses in RCTs. The subsequent chapter will then describe a systematic review to evaluate subgroup analyses performed in RCTs of therapist delivered interventions for non-specific LBP.

Chapter 3

Subgroup Analyses in Clinical Trials

3.1 Introduction

The previous chapter described how low back pain (LBP) is a very common and costly problem that affects most adults at some point during their life. It was also highlighted that identifying patient subgroups that most benefit from treatment is regarded as a high research priority in the LBP community. However, there are many arguments for and against performing subgroup analyses in general (41, 50). Some arguments against subgroup analyses are from statisticians who warn of the dangers in relation to the statistical issues associated with such analyses. On the other hand, clinicians argue that applying the overall result of a trial to individuals without considering the factors associated with an individual's response can also be dangerous. Large amounts of money are expended in conducting trials with interest in testing primary hypotheses. Therefore a simplistic argument would be to fully utilize the data to extract as much information as possible by performing secondary analyses instead of performing just the primary analyses alone.

Despite the arguments against subgroup analyses, many still see the advantages and still perform them. In fact, there are statisticians, clinicians and health professionals in academia and the pharmaceutical industry that have a genuine interest in subgroup analyses. Generally speaking, subgroup analyses that are performed are exploratory in nature; i.e. they are not pre-planned and do not aim to prove an effect. Exploratory subgroup analyses can be classified into three types. The first type is to evaluate the internal consistency of the overall results from a trial. Many areas of research conduct trials to test hypotheses regarding the overall mean treatment effect in the population of interest. Though there might be an overall effect, one cannot make the assumption that this effect is the same across subgroups. Hence conclusions assuming the effects are the same in subgroups without investigating the internal consistency might be considered a wrong thing to do. The second type of exploratory subgroup analyses are post-hoc analyses to identify subgroups with improved treatment effect. Identifying such subgroups can help improve individualized patient care by targeting treatment accordingly. Finally, the third type is to find positive inference from a trial that failed in terms of statistical significance in the primary endpoint. Although the overall mean treatment effect was not beneficial to the study population, there may be subgroups that do actually benefit from treatment.

From the three types of exploratory analyses described above, it is clear that over and above the statistical issues, there are several benefits and uses of performing subgroup analyses. The second of the three types is considered a research priority in the area of LBP research. Several subgroup analyses have been performed in the LBP literature over the years. It is therefore of interest to review the quality, conduct and reporting of these secondary analyses. Prior to doing so, it is important to understand the concept of subgroup analyses, the associated issues and review what the current proposed guidelines or recommendations are for performing subgroup analyses in RCTs in

general. Therefore, this chapter will initially describe what subgroup analysis is and the associated statistical issues without considering anything in relation to the specific clinical aspects. To better deal with the associated issues, a number of subgroup analyses guidelines have been proposed. Thus, some of the key recommendations from proposed guidelines within the context of RCTs will be described. Some concerns regarding the current recommendations will also be discussed thereafter.

3.2 Subgroup analysis concept

A randomized controlled trial (RCT) is considered the most valid way of evaluating the effectiveness of a new health care intervention (51-53). Consider a RCT that is designed to assess the effectiveness of a new intervention compared to a control arm or an existing intervention where we want to test the null hypothesis that both the interventions are equally effective in terms of some primary endpoint. Subsequently, we want to perform secondary subgroup analyses with the aim of trying to identify subgroups of patients for whom the new intervention works differently i.e. is more (or less) beneficial. Though this may seem rather easy and straightforward, in reality it is not that simple. Subgroup analyses are associated with a number of issues (discussed in this chapter) which require due consideration when performing analyses. If these issues are ignored and not appropriately acknowledged or addressed, it could lead to the over-interpretation of results which in turn might wrongly influence and misguide future research (41, 43, 44, 50, 54, 55). The statistical challenges associated with performing subgroup analyses are well documented in the literature. These challenges will be described and discussed below to familiarize the reader with the problems faced when performing subgroup analyses.

3.3 What is a subgroup analysis?

A subgroup analysis is defined as the evaluation and comparison of treatment effects in subgroups of patients defined by baseline characteristics (44). It is most commonly referred to as subgroup analysis but is also, and less commonly, referred to as moderator analysis, subset analysis and test for treatment-by-covariate interaction. For example, in a low back pain (LBP) trial comparing a new exercise regimen (intervention arm) to advice to remain physically active (control arm), the investigator may wish to test the null hypothesis that both treatments are equally effective. They then might go on to test a secondary null hypothesis, pre-specified in the study protocol, that the treatment effect does not differ (is not more or less beneficial) in the employed subgroup compared to the unemployed subgroup and subsequently test this on completion of the trial.

Any baseline characteristics that show evidence of a significant subgroup effect are also referred to as moderators of treatment effect. At this point it is important to distinguish between moderators, mediators and predictors. Moderators are baseline characteristics, measured before receiving any treatment, that have an interactive effect with treatment on outcome. Mediators are potential mechanisms, measured during treatment, that have interactive effect with treatment on outcome (56).

Predictors, as the name suggests, are variables or factors that help predict or forecast the value of a future outcome irrespective of treatment. In a statistical model, there is a clear distinction between factors that have an overall effect and factors that have a differential or moderating effect i.e. an interaction with treatment. In this PhD, the focus will be on identifying those factors that have an interaction effect with treatment.

3.4 Statistical concepts and issues with subgroup analyses

Testing for subgroup effects

Let us continue with the same example from above of a trial comparing a new exercise regimen (intervention arm) to advice to remain physically active (control arm).

Assuming the trial reaches completion; there are clearly two possible scenarios that can occur. The first scenario is when the overall outcome of the trial fails to provide evidence of a statistically significant difference between the two arms. Even though the trial failed to detect an overall difference there is still a possibility that particular subgroups of patients did benefit from the new intervention e.g. the employed subgroup. In the second scenario, the overall result of the trial concludes superiority of the intervention arm, in which case the size of the benefit might differ across certain subgroups e.g. increased benefit in the employed subgroup compared to unemployed subgroup. Observing varying treatment effects across different subgroups is also referred to as treatment effect heterogeneity.

When performing subgroup analyses, the investigator may be naturally inclined to test within individual subgroups separately. For example, the investigator may test for a treatment effect in the employed subgroup and find evidence of a significant treatment effect and then they may test for treatment effect in the unemployed subgroup and fail to detect a treatment effect. They might then erroneously conclude from this, ignoring the uncertainty, that a subgroup effect exists and that employed people do better with treatment than unemployed people. However, the uncertainty must never be overlooked. Separate tests for treatment effect only assess if the effects are different from zero in each particular subgroup; it does not test if the difference in treatment effect between both groups is different from zero. Thus, performing tests for treatment effect within subgroups separately is an incorrect approach to subgroup analyses for two reasons (41, 43, 50, 55, 57, 58). Firstly, applying this kind of an approach does not

directly evaluate the treatment effect heterogeneity of the subgroups. The most appropriate method of directly evaluating the treatment effect heterogeneity of subgroups is by using a statistical test for interaction in the regression model (54, 59, 60). The second reason is that it inflates the overall type I error rate (probability of a false positive result) due to twice as many tests being performed (multiple testing), compared to using just the interaction test. The issue of type I error inflation will be explained in more detail in section 3.5.

Table 3.1 – Summary table of treatment group by employment status interaction

	Control ($Trt = 0$)	Intervention ($Trt = 1$)
Unemployed ($X_1 = 0$)	$\mu_{00} = \beta_0$	$\mu_{01} = \beta_0 + \beta_1$
Employed ($X_1 = 1$)	$\mu_{10} = \beta_0 + \beta_2$	$\mu_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Interaction test example

Let's consider a simple regression model where Y is an outcome variable and Trt and X_1 are two independent binary variables. In this example, let's assume that Trt is the treatment indicator (control=0, intervention=1) and X_1 is an indicator of membership for the subgroup of interest, which in this example is employment status (unemployed=0, employed=1). A test for interaction can be performed to estimate the interaction effect between Trt and X_1 using a regression model and including an interaction term:

$$Y = \beta_0 + \beta_1 Trt + \beta_2 X_1 + \beta_3 Trt \cdot X_1 + \varepsilon . \quad (3.1)$$

where ε is the normally distributed error with mean 0. The interaction term in this model to be explored by the investigator is created by simply multiplying Trt and X_1 .

The term β_0 is the intercept of the line, and β_1 and β_2 are the coefficients for the treatment indicator (Trt) and employment status indicator (X_1) respectively. The coefficients of the binary variables, Trt and X_1 , are interpreted as the mean difference between the two levels being compared e.g. $Trt = 1$ compared to $Trt = 0$. The β_3 coefficient is interpreted as the interaction effect estimate i.e. the difference in the treatment effect between the employed and unemployed subgroups. To better understand the interaction effect, we can draw a 2x2 table of Trt and X_1 with the expected values (See Table 3.1). The expected values are denoted by μ_{00} and μ_{01} for the unemployed subgroup ($X_1 = 0$) who are in the control group ($Trt = 0$) and intervention group ($Trt = 1$) respectively. Similarly, the expected values are denoted by μ_{10} and μ_{11} for the employed subgroup ($X_1 = 1$) who are in the control group and intervention group respectively. To investigate if employment status moderates treatment effect, we need to see if there is a difference in the within subgroup treatment effect between the two subgroups i.e. to see if there is an interaction effect. This means, using table 3.1, that we want to evaluate:

$$(\mu_{01} - \mu_{00}) - (\mu_{11} - \mu_{10}) = \mu_{01} - \mu_{00} - \mu_{11} + \mu_{10}. \quad (3.2)$$

and test to see if it is zero. The interaction effect can be easily shown by using the regression model (3.1) to evaluate (3.2). Using (3.1), the treatment effect in the unemployed subgroup is:

$$(\beta_0 + \beta_1 + 0 + 0) - (\beta_0 + 0 + 0 + 0) = \beta_0 + \beta_1 - \beta_0 = \beta_1.$$

Similarly, the treatment effect in the employed subgroup is:

$$(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2 + 0 + 0) = \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 + \beta_2 = \beta_1 + \beta_3.$$

So β_1 is the treatment effect of those who are unemployed ($X_1 = 0$) and the treatment effect of those who are employed ($X_1 = 1$) is $\beta_1 + \beta_3$. Thus using (3.2), the difference

between the subgroups is $(\beta_1 + \beta_3) - (\beta_1) = \beta_3$. Thus β_3 is the estimate of the interaction effect that is to be tested.

The test for interaction effect uses a t-statistic to test the null hypothesis $H_0: \beta_3=0$, i.e. the interaction effect is equal to zero. The alternate hypothesis is therefore $H_1: \beta_3 \neq 0$. The t-statistic is computed by dividing the $\hat{\beta}_3$ interaction effect estimate by its estimated standard error. The t-statistic is then used to test the interaction effect by computing the associated p-value using a t-distribution and comparing it to a pre-specified significance level (e.g. 0.05). If the test for the interaction term is significant i.e. there is evidence of an interaction effect present, then we cannot interpret Trt in terms of our dependent variable Y alone as there is some dependency on the variable X_1 . In other words, the magnitude of the effect of the treatment variable (Trt) on the dependent variable Y is moderated by the subgroup variable (X_1). However it is important to note that although the interaction term is significant, this does not necessarily mean that it is a genuine effect. It might just be a spurious finding due to other reasons (discussed later on).

Types of interaction

All interaction effects can be distinctly classified as being either quantitative or qualitative (43, 44, 50). When the intervention arm is superior or inferior to the control arm in both subgroups (e.g. employed and unemployed) but varies by different degrees, then the interaction is called quantitative (Figure 3.1). On the other hand when the intervention arm is superior in one subgroup (e.g. employed) but is found to be worse in the other subgroup (e.g. unemployed) then the interaction is said to be qualitative (Figure 3.2). It is probably worth noting here that quantitative interactions can be removed by a monotonic transformation of the measurement scale whereas qualitative interactions cannot.

Figure 3.1 – Quantitative Interaction (treatment effect in the same direction)

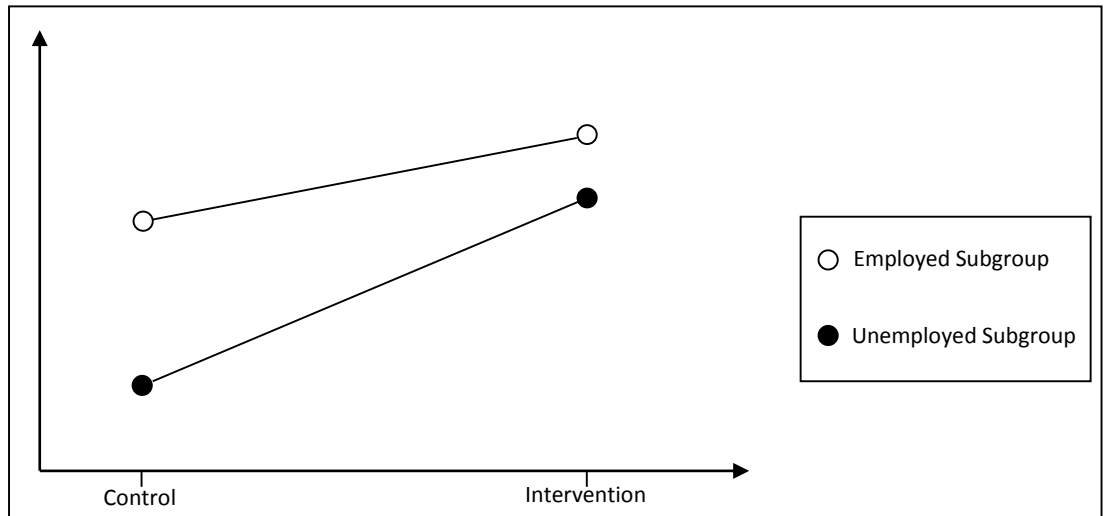
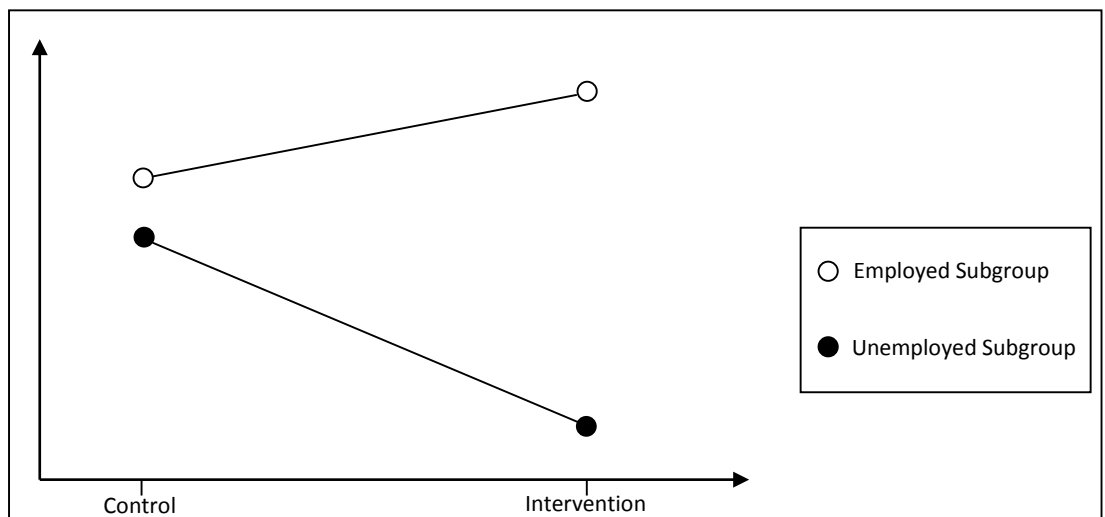


Figure 3.2 – Qualitative Interaction (treatment effect in different directions)



Multiple comparisons

When an investigator performs a single interaction test to assess treatment heterogeneity across subgroups, they must initially specify a significance level. A significance level is a fixed probability of finding a false positive result i.e. falsely rejecting the null hypothesis of equal treatment efficacy if it is actually true, also called the type I error. The significance level is kept limited to reduce the chances of the investigator making a false claim with the aim of producing a result that is believable.

The most commonly used significance value is $\alpha=0.05$ which means, in frequentist terms, that there is a 5% chance of a significant effect when the null hypothesis is true.

Most of the time there are several subgroups that are of interest either due to some clinical reasoning, evidence in previous literature or purely just for exploratory purposes. Therefore implementing the interaction test for each of these subgroups individually will inflate the overall type I error rate hence increasing the chances of a spurious finding. Let α be the significance level, where α = Type I error i.e. the probability of finding a false positive result for a single test. Then the probability of correctly failing to reject the null hypothesis for a single independent test is simply $(1 - \alpha)$. If there are m independent tests performed then the combined probability of correctly failing to reject any null hypothesis is equal to:

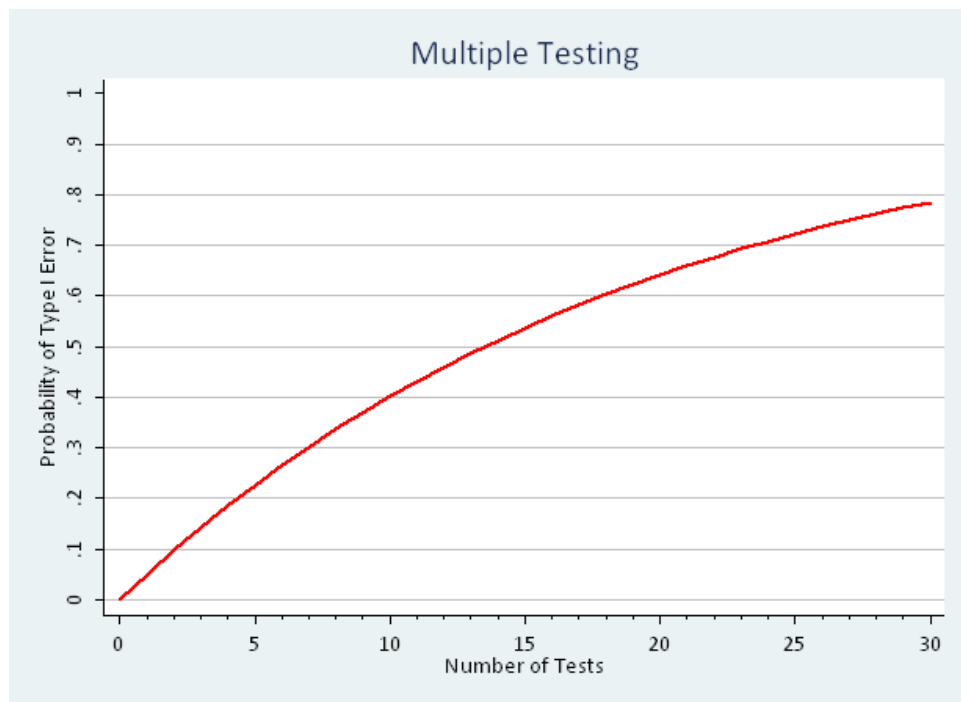
$$(1 - \alpha) \times (1 - \alpha) \times \dots \times (1 - \alpha) = (1 - \alpha)^m = \prod_{i=1}^m \{1 - \alpha\}$$

Therefore the probability that at least one of the m independent null hypotheses is falsely rejected is equal to:

$$1 - (1 - \alpha)^m = \alpha_m \quad (3.3)$$

Hence as the number of independent interaction tests (m) increases, the overall type I error rate (α_m) increases, as illustrated in figure 3.3. For example, when 5 independent interaction tests are performed ($m=5$) using a significance level of $\alpha=0.05$ then the probability, calculated using equation (3.3), of detecting at least one false positive result i.e. falsely detecting at least one statistically significant interaction out of the five, is approximately 23%. Multiple comparisons may also occur when a significant interaction has been found for a categorical variable, which therefore requires multiple tests to identify the specific category or categories that are heterogeneous.

Figure 3.3 – Probability of falsely detecting at least one significant interaction at the 5% significance level for independent hypothesis tests



There are examples of the misuse (multiple testing) and benefits of performing subgroup analyses. An overly enthusiastic investigator may inappropriately perform several subgroup analyses hence increasing the occurrence of spurious findings. The authors of the ISIS-2 study decided to demonstrate the multiplicity issue by investigating the star signs of patients in the study. The results suggested that the effects of aspirin and streptokinase were significantly different only for patients with the star signs Gemini and Libra (61). However, if conducted appropriately, subgroup analyses can be of great benefit. For example, subgroup analyses were performed in the BARI trial that compared bypass-graft surgery to angioplasty for revascularisation of patients with coronary artery disease. They found that there was a significant increase in the mortality of patients taking medication for diabetes who underwent angioplasty compared to those who underwent bypass-graft surgery (62). There was no difference in mortality for the non-diabetic patients.

When performing multiple tests, a multiplicity correction can be applied to control for the inflated overall type I error (63, 64). A multiplicity correction basically adjusts the significance level of each individual test downwards to ensure that the overall significance level remains at the same pre-specified level e.g. 0.05. If we assume the tests to be independent, we can simply re-arrange equation (3.3) and solve for α in terms of α_m to give:

$$\alpha = 1 - (1 - \alpha_m)^{\frac{1}{m}}.$$

This is often referred to as the Sidak correction which computes the α level required for each independent test to control α_m . However, the Sidak correction makes the assumption of independence whereas in reality there might be some dependency between the tests. If the tests are positively correlated then the Sidak correction controls the overall type I error rate, but it does not when there is negative correlation. An alternative method that can be used and does not assume independence is the Bonferroni correction which is formulated using Boole's inequality. Let E_i ($i=1,..,m$) denote the event that the i -th test rejects the null hypothesis when it is true i.e. the event of a false positive. Boole's inequality states:

$$P(E_1 \cup \dots \cup E_m) \leq \sum_{i=1}^m P(E_i).$$

If we specify the significance level for each test to be α , i.e. $P(E_i) = \alpha$, then Boole's inequality can be written as:

$$P(E_1 \cup \dots \cup E_m) \leq \sum_{i=1}^m P(E_i) = \sum_{i=1}^m \alpha = m\alpha.$$

This basically means that the type I error is at most $m\alpha$. Therefore, if we want to control the overall type I error at a pre-specified α_m level, we need to adjust the significance level (α) for each individual test by dividing by the total number of tests m . Thus, the Bonferroni correction is given by $\left(\frac{\alpha}{m}\right)$. When comparing the Bonferroni

correction to the Sidak correction, the Bonferroni is more conservative i.e. $\left(\frac{\alpha}{m}\right) \leq 1 - (1 - \alpha)^{\frac{1}{m}}$ for $m > 1$, as it does not assume independence.

Statistical power

The significance level and power to detect a desired main effect on a primary endpoint is established during the design stages of all trials when estimating the required overall sample size. One of the factors thereafter that can negatively influence the power of a study to test the null hypothesis is a reduction in sample size. Therefore, performing secondary analyses of subsamples will considerably reduce the power to detect an effect size of the same magnitude as that of the main treatment effect. Thus, an interaction test will only allow substantially larger effect sizes to be detected; however, in reality we would probably expect them to be smaller than the main effects.

Consider a two-arm trial with n participants in each arm that is powered to detect a main effect $\hat{\mu}_1 - \hat{\mu}_2$. We know that both $var(\hat{\mu}_1)$ and $var(\hat{\mu}_2)$ can be computed by $\frac{\sigma^2}{n}$. Thus it can be shown assuming independence that the variance of the main effect is simply

$$var(\hat{\mu}_1 - \hat{\mu}_2) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}.$$

Now consider the subgroup example in table 3.1 looking at the difference in treatment effect across employment status. This difference is evaluated by use of an interaction test $(\hat{\mu}_{01} - \hat{\mu}_{00}) - (\hat{\mu}_{11} - \hat{\mu}_{10})$. If we assume that the number of participants in each of the four subgroups is $\frac{n}{2}$, then the variance of the interaction effect equates to

$$\begin{aligned} var((\hat{\mu}_{01} - \hat{\mu}_{00}) - (\hat{\mu}_{11} - \hat{\mu}_{10})) &= var(\hat{\mu}_{01} - \hat{\mu}_{00}) + var(\hat{\mu}_{11} - \hat{\mu}_{10}) \\ &= \frac{4\sigma^2}{n} + \frac{4\sigma^2}{n} = \frac{8\sigma^2}{n} = 4 \times \left(\frac{2\sigma^2}{n}\right). \end{aligned}$$

Thus from the above, in order to observe an interaction effect that is of the same magnitude as that of the main effect size i.e. have the same power, the original sample size would have to be increased by four-fold and even more if the subgroups are not of equal size (58, 59). However, taking such an approach to ensure subgroup analyses are powered will have big implications on the trial as a whole. Such a drastic inflation of the sample size would prolong the overall duration of the trial, and as a direct consequence would incur increased costs to run the trial. Thus such an approach to power subgroup analyses is extremely costly, very time consuming and probably not worth pursuing.

An alternative and tempting option to increase power when testing interaction effects would be to make the significance level less stringent (e.g. $\alpha=0.20$) making it easier to identify subgroup effects. However, a less stringent significance level means that there is an increase in the probability of a false positive finding i.e. falsely rejecting the null hypothesis when it is true. Moreover, it has been shown that increasing the significance level for interaction tests does not always usefully raise the power to a level that is considered adequate and is therefore not advised (65, 66). If increasing the significance level does not inflate the power to an acceptable level e.g. 80%, then this option is probably not worth pursuing as it would just increase the chances of detecting spurious findings.

3.5 Subgroup analyses guidance

The key statistical concepts and issues associated with subgroup analyses were detailed in section 3.4. These issues are very well recognized in the literature and as a result, a number of subgroup analyses guidelines have been proposed (41, 43, 44, 67). Although the statistical issues of power and multiplicity remain, complying with these

guidelines maximises the credibility of the results. The key recommendations from these guidelines will now be described.

Interaction test

The most appropriate method of directly evaluating the treatment effect heterogeneity of subgroups is by using a statistical test for interaction in the regression model (54, 59, 60).

Pre-planned and post hoc analyses

Having conducted a trial, it can be extremely tempting to look at the data and perform several analyses to test for treatment heterogeneity within subgroups of participants. This is considered a misuse of data and is sometimes referred to as data dredging. Such an approach is inappropriate as it increases the chances of false positive findings, as explained earlier, because the number of such analyses that could be done is very large. Therefore an important recommendation to avoid this is to pre-specify the actual subgroup analyses to be undertaken, preferably in the study protocol i.e. specified without looking at the data, for them to be classed as a priori analyses. However, it might well be that an investigator genuinely didn't think about looking for subgroup effects until after (post-hoc) the study. Alternatively, they may have pre-specified a few subgroups of interest but then felt the urge to explore other subgroups post-hoc as well. Whatever the true scenario is, it is recommended that a clear distinction be made between pre-specified analyses and post-hoc analyses. Pre-specified analyses are for testing hypothesis (confirmatory analyses) whereas post-hoc analyses (exploratory analyses) are for generating hypotheses to be tested in future studies (68-70).

It is important to note here that pre-specifying subgroup analyses to be undertaken does not resolve the issue of multiple testing and the inflated type I error rate. For this

reason, it is advised that the number of pre-specified analyses be kept to a minimum and that the results be interpreted with caution. When pre-specifying a limited number of subgroup analyses, it is recommended that the following information be provided:

- Specification of a limited number of baseline characteristics to be investigated.
- Specification of the cut point to be used for any continuous baseline variables, including some justification of the chosen cut point based on previous literature. If no specific cut point is documented in the literature, it is suggested that the median is used.
- A clear explanation as to what is guiding the subgroup analyses with reference to relevant literature and findings from previous studies.
- The end point of interest should be clearly specified.
- The statistical method to be used to test for subgroup effects (interaction test) should be specified.
- A statement, with some justification, to indicate the expected subgroup effect size and direction.

There is a possibility that during the course of the trial and before any data are analyzed, some new information may emerge in the literature that was not readily available during the design stages of the study. In such instances an amendment to the study protocol will suffice.

Multiplicity adjustment

Subgroup analyses guidelines recommend that an appropriate multiplicity correction (e.g. Bonferroni or Sidak) should only be applied when testing a small number of pre-specified hypotheses i.e. performing confirmatory analyses (41). Multiplicity

corrections are not required when performing post-hoc exploratory analyses since the aim is to simply generate hypotheses rather than test them (41).

Reporting

All subgroup analyses that are conducted should be reported. It is recommended that the interaction effect estimate, 95% confidence interval and p-value are presented when reporting subgroup analyses. Due to the issue of multiple testing as described above, it is strongly recommended that the results of any subgroup analyses be interpreted with caution (43, 59).

3.6 Issues with current recommendations

The concepts, issues and current recommendations in proposed guidelines for performing subgroup analyses have thus far been explained in this chapter. However there are a number of issues with these recommendations. One issue that arises is the selection of subgroups to be investigated. It is recommended that a limited number of subgroups be pre-specified a-priori, preferably in the study protocol, with clear justification. This depends on findings from previous relevant literature or clinical justification and can thus be a very subjective process. Although some subgroups may be identified during this process, it is possible that important subgroup effects in other baseline covariates that have not been pre-specified may go unnoticed. This could be due to two reasons; either because they have not been previously investigated or that the investigator failed to report particular results as the treatment showed signs of being more harmful than good. Therefore, one could argue that since conventional subgroup analyses are exploratory in nature, all possible baseline covariates should be investigated to ensure no important subgroup effects are missed instead of just investigating pre-specified subgroups only. Moreover, it is recommended that the cut-point selection for continuous baseline covariates be pre-specified based on cut-points

used in previous literature, if available. This again is a subjective process that could miss important subgroup effects of subgroups defined using another cut-point. Hence, one could argue that either numerous cut-points or all possible cut-points should be investigated when performing subgroup analyses.

Another issue with conventional subgroup analyses is that the selected subgroups are assessed independently using separate interaction tests. Testing for subgroup effects separately might not be the best way to identify subgroups as we know that each individual can be described by several baseline characteristics. Moreover, one may be easily inclined to incorrectly interpret the results of separate interaction tests using a statement combining the subgroups. For example, let's say that a subgroup analysis using separate interaction tests finds both age and sex to have significant interaction effects; more precisely that younger participants benefit more from treatment and females benefit more from treatment. These are two separate subgroups that have been identified and so should be interpreted separately; however, one may falsely interpret this by saying younger females benefit more from treatment i.e. combining them into one subgroup. This statement might not necessarily be true since the single particular subgroup defined by the two characteristics (age and sex) was not tested. In reality, it is highly likely that there are several patient characteristics that are involved in determining a subgroup effect and not individual characteristics. This therefore suggests that alternative statistical methods need to be sought to perform subgroup analyses to incorporate multiple baseline characteristics. More specifically, the method should have some systematic way of performing interaction tests to identify subgroups whilst simultaneously forming a subgroup defined by several characteristics.

3.7 Discussion

The major issues involved with performing subgroup analyses are well recognised and documented in the literature. There is some debate surrounding the use and worth of subgroup analyses in studies, most of which is subjective. Despite the issues involved, the general consensus is that it is highly beneficial to perform subgroup analyses to generate important hypotheses that can be tested in future research and provide additional information for clinical guidelines. Clearly the ISIS-2 and BARI studies exemplify two extremes of the outcomes from subgroup analyses. It is therefore essential that the planning, analyses, reporting and interpretation of subgroup analyses conducted using interaction tests be performed in a rigorous methodological framework as proposed by a number of guidelines (41, 44). Despite the analyses lacking power, following recommendations will at least ensure that they are of a good credible standard, thus strengthening the implications for future research.

While it has been recommended that treatment effect heterogeneity may be formally assessed using a statistical test for interaction, there are key issues that still remain, as described in this chapter. The key issues are a lack of power to detect genuine treatment-covariate interactions and also the issue of multiplicity. One method of overcoming the lack of power is to power the study to incorporate subgroup analyses. However, this would require an extremely large study that would be very costly and time consuming. An alternative option would be to collect a repository of individual patient data from several trials of similar interventions that have already been conducted and then perform subgroup analyses of the pooled dataset. While this may improve the lack of power issue, the issue of multiplicity still remains due to the parametric nature of the statistical methods used and it also introduces other challenges.

Although following current proposed guidelines ensures the credibility of subgroup analyses, the issues of power, multiplicity, failure to investigate other potentially important subgroups and incorporating multiple patient characteristics still remains. Meta-analyses or individual patient data meta-analyses can be used to improve the issue of power. Of the two, individual patient data meta-analysis has the advantage of allowing patient-level covariates to be fully explored for subgroup effects (71). While the issue of power can be substantially improved by using individual patient data from several similar trials, the issue of multiplicity still exists if current parametric statistical methods for subgroup analyses are applied. This therefore highlights the need for the development of a non-parametric or data driven approach to performing subgroup analyses in an individual patient data meta-analyses context focusing on subgroup identification using several patient baseline characteristics.

This chapter reviewed the recommendations of current proposed guidelines for performing subgroup analyses in trials in general. In non-specific LBP trials, several papers have performed and published their findings from subgroup analyses. Therefore, it is of interest to evaluate the quality, conduct and reporting of subgroup analyses performed in the LBP literature in accordance with current guidelines. Even though a number of issues with current recommendations were highlighted in this chapter, it is still important to review the LBP literature in line with the current guidelines. Thus, the following chapter will describe a systematic review of subgroup analyses performed in trials of therapist delivered interventions for the management of non-specific LBP. The subsequent chapter (Chapter 5) will then go on to describe alternative statistical approaches for performing subgroup analyses in order to address the issues highlighted with regard to the current guidelines.

Chapter 4

Systematic review of subgroup analyses in non-specific low back pain trials

4.1 Introduction

The previous chapter provided a detailed description of the statistical concepts and issues associated with subgroup analyses. To better control for these issues, several proposed guidelines have been published to ensure, if used, that subgroup analyses are of a good standard. Thus the previous chapter also presented the key recommendations from proposed guidelines for performing subgroup analyses. A discussion regarding several issues with these guidelines and conventional subgroup analyses was presented thereafter. Despite these issues, it is still important to review the current LBP subgroup analyses literature in line with current guidelines. As mentioned in chapter 2, the overall quality of subgroup analyses in the area of LBP is unknown. Therefore, this chapter will report a systematic review of the quality,

conduct and reporting of subgroup analyses performed in the area of LBP to allow interpretation of their findings. The content of this systematic review has been published in the Spine Journal (see appendix A) (1).

Objective

The objective of this systematic review is to identify RCTs of therapist delivered interventions for non-specific LBP that have performed secondary analyses in the form of subgroup analyses and then evaluate the quality of these analyses. Furthermore, the conduct and reporting of subgroup analyses will also be assessed.

4.2 Methods

Search Strategy

We searched MEDLINE (1948 to July 2013), Ovid MEDLINE(R) In-process & Other Non-Indexed Citations, Embase (1974 to July 2013), Web of Science and Citation Index and Cochrane Controlled Trials Register (CENTRAL). We sought to identify all papers reporting RCTs of therapist delivered interventions for non-specific LBP that performed subgroup analyses. For that reason, we initially searched for 'low back pain' terms, 'RCT' terms and key subgroup analyses terms such as 'subgroup', 'effect modifier' and 'moderator'. These search terms only identified papers that had subgroup analyses terms in the title or abstract and therefore missed out publications that had these terms in the main text only. We therefore re-ran our searches to identify all RCTs of therapist delivered interventions in the area of non-specific LBP using only the keywords for 'low back pain' and 'RCTs'. A full list of the search terms used can be found in appendix B.

Selection of papers

We scanned all of the titles and abstracts from the search results retrieved to identify all papers potentially reporting a subgroup analysis of an RCT testing a therapist delivered intervention for LBP. Here we define a therapist as a person trained in administering any of the available recommended treatments, excluding drug interventions and surgical interventions, for the management of LBP. We then examined the full text of every paper to see if they performed some form of subgroup analyses and then using specific inclusion and exclusion criteria we decided which papers to include in the final review (Table 4.1). Papers testing a clinical prediction rule were excluded because they only test to see if a pre-defined rule works rather than performing tests to identify subgroup characteristics that modify treatment effect. Papers interested in treatment effect modification over time were also excluded because typically the primary analysis in a LBP study evaluates the effect at a specific time-point e.g. at 12 months, and not over time. Although studies do report outcomes at multiple time-points, it is the primary analysis endpoint that is of interest.

Table 4.1 – Inclusion and exclusion criteria used to select papers

Inclusion criteria:
<ul style="list-style-type: none">• Randomized controlled trials• Published in English-language• Participants aged 18 years or more with history of non-specific LBP (including sciatica)• Therapist delivered interventions for non-specific LBP (including psychological interventions and intensive rehabilitation programmes)• Primary or secondary analysis of RCTs reporting that a subgroup analysis had been conducted
Exclusion criteria:
<ul style="list-style-type: none">• LBP with known likely cause (fracture, infection, malignancy specific cause, ankylosing spondylitis and other inflammatory disorders)• Studies investigating disorders additional to non-specific LBP e.g. non-specific LBP and neck pain• Outcome not a valid clinical measure of non-specific LBP e.g. number of day's sick leave• Testing a clinical prediction rule• Treatment effect modification over time i.e. treatment x moderator x time• Pooled datasets of similar trials

Quality of subgroup analysis

Papers included in the final review were assessed for the quality of subgroup analyses using the following criteria as proposed by Pincus et al (72).

- 1) Was the subgroup analysis specified a priori?
- 2) Was the selection of subgroup factors for analysis theory/evidence driven?
- 3) Were subgroup factors measured prior to randomization?
- 4) Was measurement of subgroup factors measured by adequate (reliable and valid) measurements, appropriate for the target population?
- 5) Does the analysis contain an explicit test of the interaction between moderator and treatment?

The quality assessment classifies papers as either providing confirmatory findings or exploratory findings. Confirmatory findings support hypotheses about moderators (hypothesis testing) whereas exploratory findings inform future research (hypothesis generating). If a paper satisfies all five criteria then its findings are regarded as being confirmatory. Papers satisfying criteria three, four and five only are regarded as having exploratory findings. All other remaining papers are regarded as having insufficient findings. The quality of subgroup analyses in all of the identified papers was also assessed separately by two independent reviewers; Dr Shilpa Patel and Dr Siew Wan Hee. Any discrepancies at the end of the process were resolved through discussion.

Assessment of Conduct and Reporting

There are a number of proposed guidelines that exist for the conduct and reporting of subgroup analyses to ensure that the conclusions drawn are plausible (41, 43). These guidelines were used to help evaluate the conduct and reporting in the papers

identified from the literature search. In particular, three areas were assessed; 1) design and methods, 2) reporting of results and 3) interpretation and discussion.

The design and methods was assessed for all papers whereas the reporting of results and the interpretation and discussion were only assessed for those papers that used interaction tests for subgroup analyses. The conduct and reporting of all papers was examined to see if they conformed to the following four key recommendations in the area of subgroup analyses (41, 43).

- Exact subgroup definitions should be given beforehand for continuous and categorical variables along with some justification to avoid post-hoc data dependent definitions of subgroups.
- Subgroup analyses should be performed on the primary outcome in the study. This is simply because trials are designed to detect differences in the primary outcome only; therefore performing subgroup analyses on any other outcome measure will substantially reduce the power.
- A differential subgroup effect should be formally evaluated using a statistical test for interaction and the interaction effect reported. Performing tests within individual subgroups and then comparing the results is an incorrect approach to performing subgroup analyses as it does not directly evaluate the subgroup effect.
- The number of subgroup analyses to be performed should be kept to a minimum. This is to avoid the issue of false-positive discovery (type-I error inflation) due to multiple testing; a well-known issue if there are several subgroups of interest. Any concerns regarding multiplicity should be acknowledged and addressed appropriately e.g. applying a Bonferroni correction.

4.3 Results

Selection of papers

We identified 4,933 papers from the screening of titles and abstracts of which 4,873 papers were excluded. The full texts for the remaining 60 papers were examined and a further 21 papers were excluded based on the inclusion and exclusion criteria. The remaining 39 papers were included in the final review (see Figure 4.1, Tables 4.2 and 4.3).

Most of the included papers were of studies conducted in the Netherlands 8(21%), UK 8(21%) or the USA 8(21%). The median study size of the included papers was 223, range 100 to 3093. Twenty-nine papers (74%) performed subgroup analyses on a total study size of around 300 or fewer; the remaining 10 (26%) papers had more than 400 patients.

Quality of subgroup analyses

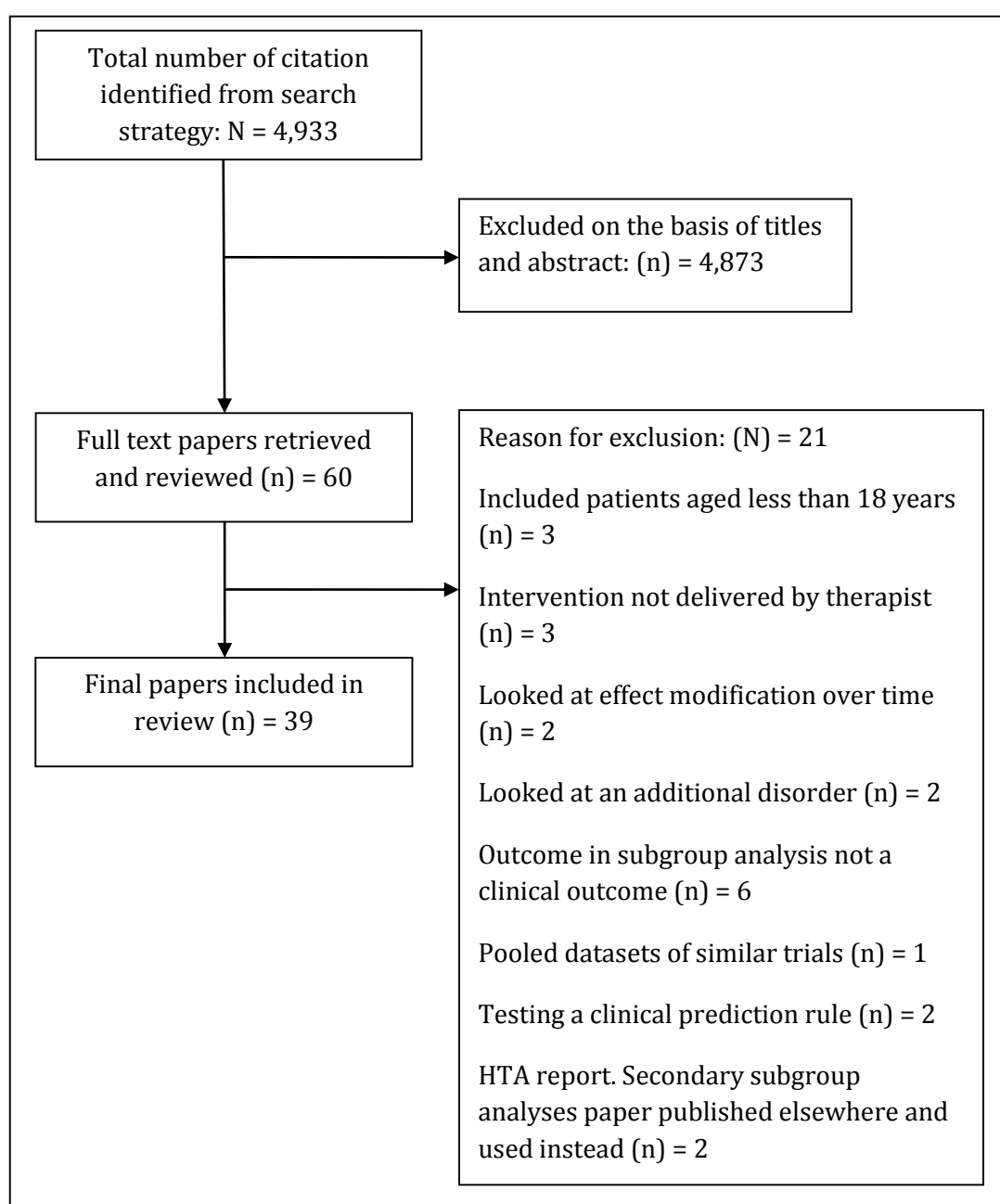
Of the 39 papers; three (8%) papers met all five criteria and therefore provided confirmatory findings (73-75), 18(46%) papers provided exploratory findings, i.e. they met criteria three, four and five, and 18(46%) papers provided insufficient findings (Table 4.2).

Assessment of conduct and reporting: Design and methods

Only one trial was designed to have adequate power to detect important treatment-covariate interactions; however no specific subgroups of interest were specified a priori (76). The majority of the papers, 31(79%), did not pre-specify which subgroups were to be investigated in the analyses. Eight papers pre-specified subgroups for confirmatory analyses (73-75, 77-81). Six of these additionally performed exploratory analyses; however this distinction was not always made clear at the outset. Baseline

characteristics that were pre-specified for subgroup analyses included age, sex, baseline Roland and Morris Disability Questionnaire (RMDQ) score, psychological distress, work load, history of back pain, radiculopathy, patient preference, catastrophizing, coping, pain self-efficacy (PSE), anxiety, depression, stress, troublesomeness, fear-avoidance, patient expectation, pain changes with position or movement, presence of leg pain, pain worse with flexion and duration (73-75, 77-81).

Figure 4.1 – Flow diagram



Four of the papers did not mention in the methods section that subgroup analyses would be performed and just included the analyses in the results section of the paper (82-85). All 39 papers, regardless of whether they performed a formal test for interaction, did measure the subgroups of interest before randomization. The majority of these subgroup factors, 33(85%), were measured using adequate (reliable and valid) measurements. The commonest subgroup factor that was not measured adequately was patient expectation.

Only one (3%) paper gave any indication as to the size and direction of the subgroup effect they were expecting prior to performing the analyses (73). Three (8%) of the papers only provided a prediction for the direction of the subgroup effect and around a third of the papers provided some justification regarding the choice of subgroups to be analysed.

Thirteen of the papers were able to provide exact definitions of subgroups although only five (13%) provided clear justification for the cut-points used to define subgroups. All 39 papers performed subgroup analyses on the primary outcome of which four (10%) also performed subgroup analyses using secondary outcomes.

Two papers in particular reported around sixty interaction tests in addition to the primary analyses; implying a substantial inflation of the overall type-I error rate, thus increasing the chance of detecting spurious findings (12, 74). Of the three papers that provided confirmatory findings, only one of them acknowledged and dealt with the issue of multiple testing. They did this by including a multiplicity correction (Bonferroni correction) for the confirmatory subgroup analyses they performed (75).

Twelve (31%) of the papers did not use a statistical test for interaction to assess for treatment effect modification. Two of these papers did not give any indication as to what statistical method they used for this (82, 86). Two papers looked at correlations between individual subgroups and outcomes within each treatment arm separately (87, 88). Two papers used t-tests between treatment groups within individual subgroups (77, 84). Five papers used either multiple linear regression or multiple logistic regression for each individual subgroup (89-93). Finally, one paper compared the medians across three trial arms within individual subgroups using Kruskal-Wallis tests (78).

Assessment of conduct and reporting: Reporting of results

There is some confusion in the papers between investigating 'subgroup effects' and investigating 'differential subgroup effects' where the former investigates a specific subset or subpopulation of the entire sample for a main effect and the latter investigates treatment effect heterogeneity using an interaction test between subgroups defined by factors measured prior to treatment (50). Twenty-seven (69%) of the papers used a statistical test for interaction to perform subgroup analyses. Four of these reported subgroup analyses within individual subgroups (94-97), ten papers reported results from interaction tests (12, 73-75, 80, 81, 98-101) and the remaining 13 papers either did not report any results at all or just reported the p-value for the interaction term (37, 76, 78, 79, 83, 85, 102-108). Six papers reported both the interaction effect sizes with confidence intervals and the corresponding p-values (12, 81, 98-101), four papers reported only the interaction effect sizes with confidence intervals (73-75, 80), eight papers reported the p-values only (37, 76, 85, 95, 97, 104, 105, 107) and nine papers did not report either the interaction effect sizes and confidence intervals or p-values (78, 79, 83, 94, 96, 102, 103, 106, 108).

Assessment of conduct and reporting: Interpretation and discussion

Four out of 27 papers that performed interaction tests reported subgroup analyses within individual subgroups and thus based the interpretations and discussion on this as well. Around a third of the papers provided supporting or contradictory findings from other relevant studies. Twelve of the twenty-seven papers that used an interaction test reported significant findings, of which only two suggested the identified subgroups for investigation in future studies. Twelve papers acknowledged the limitations of performing subgroup analyses in the discussion section of the paper.

4.4 Discussion

The objective of this systematic review was to assess the quality, conduct and reporting of subgroup analyses performed in RCTs of therapist delivered interventions for the management of non-specific LBP. This quite possibly is the first study of the overall quality, conduct and reporting of subgroup analyses in the area of low back pain.

Reporting quality

Many authors have performed subgroup analyses or have attempted some form of subgroup analyses. There is some clear confusion between investigating 'subgroup effects' and investigating 'differential subgroup effects' (50). The results of the quality assessment suggest that only three of the papers are able to provide confirmatory findings and that the majority of the papers provide exploratory or insufficient findings. These results are solely based on the outcomes from the quality assessment; the results of an apparently high quality subgroup analysis need to be interpreted in light of the quality of the main study. The general content and reporting of these papers in relation to subgroup analyses i.e. in terms of design and methods, results, interpretation and discussion, is quite poor. These papers can be seen as missed opportunities. Several of these subgroup analysis papers could have either used more

appropriate methodology i.e. statistical test for interaction, improved the standard of reporting or both. Had this been done, they would have been able to contribute valuable information to the existing pool of subgroup related literature in the area of low back pain. That nearly half of the identified papers provided insufficient findings raises concerns that the already published subgroup literature be misinterpreted when considering future subgroups research or making treatment choice. Although most subgroup analyses lack power and are of an exploratory nature, a well conducted and reported subgroup analysis will ensure credibility of findings that can be tested in future studies.

Sample size

All but one of the papers that were reviewed had inadequate sample size and were thus substantially underpowered to detect any meaningful interaction effects in the primary outcome. The one trial designed and powered to detect important treatment-covariate interactions did not, however, pre-specify subgroups of interest (76). Some subgroups were found to significantly moderate treatment effect in this trial; however it was quite disappointing to see only p-values reported. As part of the systematic review for the main funded project, we did contact the author to provide us with more information but they were unable to do so.

Lack of power is a well-known issue associated with subgroup analyses. A simple model proposed by Lachenbruch, assuming equal group sizes, suggests that a total sample size of approximately 503 participants provides 80% power to detect a moderate standardized effect size of 0.5 in the interaction effect at a two-sided 5% significance level between two arms where the outcome is continuous and normally distributed (109). To date, we are only aware of four trials with subgroup analyses that have this sample size ($n \geq 500$) (34, 36, 76, 103). These were high-quality RCTs that

were designed and powered to detect a standardized mean difference of around 0.4 in the main effect. However inspection of the main effect sizes suggest that the standardized effects were much lower and only ranged from 0.12 to 0.23 (37). If the larger trials are failing to pick up a moderate standardized main effect size, it is highly unlikely that any plausible and clinically meaningful interaction effects will be detected in any subgroup analyses unless the interaction effect is considerably larger than the main effect. In a simple model, with equal numbers in each subgroup such a large interaction effect with a small main effect of treatment would consequentially mean that there is a large benefit in one sub-group and a smaller harm in the other subgroup.

A trial to identify a differential subgroup effect needs to be approximately four times larger than a trial powered to detect a main effect of the same magnitude only (59). Any such trial would only be able to test a moderator of treatment effect for one subgroup. Unless there is an overwhelming a priori hypothesis that needs testing in such a study, this is unlikely to be a worthwhile expenditure of academic effort and funders' resources. We are not aware of any such overwhelming a priori hypothesis. Furthermore, our existing pool of baseline predictors only explain about a third of the variance in outcome making it unlikely that we can identify a single strong moderator of treatment effect to underpin such a strong a priori hypothesis (110). Even if such a study was designed it would, if it was to inform clinical practice, need to ensure that whatever moderators were proposed could easily be applied in clinical practice. Therefore, it is clear that different approaches are needed.

One possible alternative approach, as mentioned in the previous chapter, would be to collect individual patient data from several similar trials. Although collecting individual patient data would improve the issue of insufficient power, the issue of multiple testing and hence the inflation of the family-wise type I error rate (probability of a false

positive result) still remains. This therefore highlights the need to identify or develop alternative statistical approaches for performing subgroup analyses. Moreover, as highlighted in the previous chapter, conventional subgroup analyses typically use interaction tests to investigate patient characteristics one at a time. However, it is quite obvious that patients have multiple characteristics that also need to be investigated either simultaneously or in some stepwise fashion. Thus this further highlights the requirement of novel statistical approaches to allow for the identification of multiple patient characteristics or clusters of moderators that would identify who is most (or least) likely to benefit.

In conclusion, finding moderators of treatment effect has been identified as a high research priority internationally for the management of non-specific LBP in a step towards better individualized patient care. The findings of this review suggest that the majority of subgroup analyses performed in low back pain trials to date are only able to provide exploratory or insufficient findings. Papers with insufficient findings are not very credible and could potentially provide false implications for guiding future research. Moreover, the general content and reporting of subgroup analyses is rather poor. It is thus recommended that authors use available guidelines when performing subgroup analyses to ensure that they are reliable and of a good standard (41, 44, 67). We do, however, have serious concerns that current approaches are inadequate for the task at hand. There is a need to find and develop alternative statistical methods for performing subgroup analyses to overcome or better deal with the existing issues associated with current methodology. The following chapter will therefore review the generic statistical methodology currently used to perform subgroup analyses.

Table 4.2 – Summary of included papers ordered by subgroup quality assessment

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up	Subgroups Identified (Interaction test only)
Confirmatory Findings	Sheets	2012	Australia	148	First-line care group vs McKenzie group	Pain measured at 1 week and 3 weeks. Global perceived effect at 3 weeks.	None
	Smeets	2009	Australia & New Zealand	259	Exercise and advice vs Exercise and sham advice vs Sham exercise and advice vs Sham exercise and sham advice	Pain intensity (11 point scale) and Patient-specific function scale (0-10 scale) measured at baseline 6 weeks and 52 weeks	None
	Underwood	2011	UK	701	Advice plus Cognitive behavioural intervention vs Advice only	RMDQ and MVK measured at baseline, 3 months, 6 months and 12 months	Age & Employment
	Becker	2008	Germany	1378	Multifaceted guideline implementation (GI) vs GI plus motivational counselling (MC) vs Postal dissemination of guideline (Control)	Hannover Functional Ability Questionnaire measured at baseline and 6 months	None
Exploratory Findings	Cecchi	2012	Italy	210	Back school vs Individual physiotherapy vs Spinal manipulation	RMDQ measured at baseline, 3 months, 6 months and 12 months	None
	Cherkin	1998	USA	321	Physical therapy vs	Bothersomeness of	Mental Health

				Chiropractic manipulation vs Educational booklet	symptoms and RMDQ measured at baseline, 4 weeks and 12 weeks	
Cherkin	2001	USA	262	Chinese acupuncture vs Therapeutic Massage vs Self-care education	Bothersomeness of symptoms and RMDQ measured at baseline, 4 weeks, 10 weeks and 1 year	None
Cherkin	2009	USA	638	Individualised acupuncture vs Standardized acupuncture vs Simulated acupuncture vs Usual care	Bothersomeness of symptoms and RMDQ measured at baseline, 8 weeks, 26 weeks and 1 year	None
Hansen	1993	Denmark	180	Intensive dynamic back-muscle exercise vs Conventional physiotherapy vs Placebo control semihot packs and light traction)	Pain level (10 point scale) measured at baseline, 4 weeks, 6 weeks and 1 year	None
Hay	2005	UK	402	Brief pain management vs Manual physiotherapy	RMDQ measured at baseline, 3 months and 12 months	None
Juni	2009	Switzerland	104	Standard care alone vs Standard care plus Spinal Manipulative Therapy (SMT)	Pain intensity (11 point scale) and analgesic use measured at baseline, days 1 to 14 and 6 months	None
Karjalainen	2004	Finland	170	Mini-intervention	Pain intensity (11 point	Perceived risk

				group vs Worksite visit group vs Usual care group	scale) measured at baseline, 3 months, 6 months, 1 year and 2 years	for not recovering & type of occupation (comparing mini-intervention vs usual care and worksite visit vs usual care)
Kole-Snijders	1999	Netherlands	159	Operant behavioural treatment with cognitive coping skills training (OPCO) vs Operant behavioural treatment with group discussion (OPDI) vs Waiting list control (WLC)	Main outcome unclear. Outcomes measured at post-treatment, 6 months and 1 year	None
Roche	2007	France	132	Active individual therapy (AIP) vs Functional restoration program (FRP)	Main outcome unclear. Outcomes measured at baseline and 5 weeks	Sorenson score
Sherman	2009	USA	638	Individualised acupuncture vs Standardized acupuncture vs Simulated acupuncture vs Usual care	Bothersomeness of symptoms and RMDQ measured at baseline, 8 weeks, 26 weeks and 1 year	Baseline RMQ
Smeets	2006	Netherlands	223	Active physical	RMDQ measured at	Baseline RMQ

				treatment (ATP) vs Cognitive behavioural treatment (CBT) vs Combined APT and CBT (CT) vs Waiting list (WL)	baseline, 10 weeks, 6 months and 12 months	
Smeets	2008	Netherlands	223	Active physical treatment (ATP) vs Graded activity with problem solving training (GAP) vs Combination treatment (CT) vs Waiting list (WL)	RMDQ measured at baseline, 10 weeks, 6 months and 12 months	None
Tilbrook	2011	UK	313	Yoga vs Usual care	RMDQ measured at baseline, 3 months, 6 months and 12 months	None
Underwood	2007	UK	1334	Control (Best care in General Practice) vs Exercise programme vs Spinal manipulation vs Combined treatment (manipulation and exercise)	RMDQ measured at baseline, 3 months and 1 year	Expectation
van der Hulst	2008	Netherlands	163	Roessingh Back Rehabilitation (RRP) vs Usual care	RMDQ measured at baseline, 1 week after treatment and 4 months after treatment	Pain intensity & Depression
Witt	2006	Germany	3093	Acupuncture vs Control (delayed	Hannover Functional Ability Questionnaire	Initial back pain, age &

					acupuncture treatment 3 months later)	(0-100 scale) measured at baseline, 3 months and 6 months	years of schooling
Insufficient Findings	Bendix	1998	Denmark	816	Functional restoration (FR) program vs Outpatients program (Control)	Main outcome unclear. Outcomes measured at baseline and 1 year	
	Beurskens	1995	Netherlands	151	Traction vs Sham traction	GPE and severity measured on visual analogue scale (VAS) at baseline and 5 weeks	
	Bishop	2011	USA	112	Supine thrust technique vs Side-lying thrust vs Non-thrust technique	ODQ measured at 1 week, 4 weeks and 6 months	None
	Carr	2005	UK	237	Group exercise programme vs Individual physiotherapy	RMDQ measured at baseline, 3 months and 6 months	
	Ferreira	2009	Australia	191	General exercise vs Motor control exercise vs Spinal manipulative therapy	GPE (11 point scale), Patient specific functional status, RMDQ, Pain intensity (10 point scale) and spinal stiffness measured at baseline and 8 weeks	None
	Glasov	2010	Australia	100	Laser acupuncture vs Sham acupuncture (control)	Pain (VAS) measured at baseline, immediately after treatment, 6 weeks and 6 months	

Gudavalli	2006	USA	235	Flexion distraction (FD) vs Active trunk exercise protocol (ATEP)	Perceived pain (VAS), RMDQ and SF-36 measured at baseline, 4 weeks, 3 months, 6 months and 1 year	
Hsieh	2004	China	146	Acupressure vs Physical therapy	Short-form pain questionnaire measured at baseline, 4 weeks and 6 months	
Jellema	2005	Netherlands	314	Minimal intervention strategy (MIS) vs Usual care	RMDQ, perceived recovery (7 point scale) and sick leave measured at baseline, 6 weeks, 13 weeks, 26 weeks and 1 year	
Johnson	2007	UK	234	Group exercise and education using a cognitive behavioural approach vs Usual care	Pain (VAS) and RMDQ measured at baseline, 3 month, 9 month and 15 months	Patient preference
Kalauokalani	2001	USA	166	Acupuncture vs Massage (Subanalysis of Cherkin 2001 paper)	RMDQ measured at baseline, 4 weeks, 10 weeks and 1 year	Patient expectations
Mellin	1989	Finland	456	Inpatient treatment vs Outpatient treatment vs Control (Advice)	Low back pain disability index (scale 0-45) measured at baseline and 3 months	
Moffett	2004	UK	187	Exercise vs Usual care	RMDQ measured at baseline, 6 weeks, 6 months and 1 year	
Myers	2008	USA	444	Usual care vs Usual	RMDQ measured at	None

				care plus patient choice of acupuncture, chiropractic or massage	baseline, 5 weeks and 12 weeks	
Seferlis	1998	Sweden	180	Manual therapy program (MTP) vs Intensive training program (ITP) vs General practitioner program (GPP)	Main outcome unclear. Outcomes measured at baseline, 1 month, 3 months and 12 months	
Thomas	2006	UK	241	Traditional acupuncture vs Usual care	Bodily pain dimension of the SF-36 (0-100 scale) measured at baseline, 3 months, 12 months and 24 months	Expectation
van der Roer	2008	Netherlands	114	Intensive group training protocol vs Guideline group	RMDQ measured at baseline, 6 weeks, 13 weeks, 26 weeks and 52 weeks	
Vollenbroek-Hutten	2004	Netherlands	163	Roessing Back Rehabilitation (RRP) vs Usual care	RMDQ measured at baseline, 1 week after treatment and 4 months after treatment	

*RMDQ - Rolland and Morris Disability Questionnaire; MVK - Modified Von Korff (pain and disability); GPE - Global perceived effect; ODQ - Oswestry Disability Questionnaire;

Table 4.3 – Summary of excluded papers

Paper	Reason for exclusion
Childs JD, Flynn TW, Fritz JM. A perspective for considering the risks and benefits of spinal manipulation in patients with low back pain. <i>Manual Therapy</i> 2006;11:316-20	Testing a clinical prediction rule (in an uncontrolled study)
Costa LO, Maher CG, Latimer J, Hodges PW, Herbert RD, Refshauge KM <i>et al.</i> Motor control exercise for chronic low back pain: a randomized placebo-controlled trial. <i>Physical Therapy</i> 2009;89:1275-86.	Look at effect modification over time
Faas A, Chavannes AW, van Eijk JT, Gubbels JW. A randomized, placebo-controlled trial of exercise therapy in patients with acute low back pain. <i>Spine</i> 1993;18:1388-95.	Included patients aged less than 18 years
Faas A, van Eijk JT, Chavannes AW, Gubbels JW. A randomized trial of exercise therapy in patients with acute low back pain. Efficacy on sickness absence. <i>Spine</i> 1995;20:941-7.	Included patients aged less than 18 years and outcome in subgroup analyses not a clinical measure of low back pain (sickness absence)
George SZ, Fritz JM, Childs JD, Brennan GP. Sex differences in predictors of outcome in selected physical therapy interventions for acute low back pain. <i>Journal of Orthopaedic & Sports Physical Therapy</i> 2006;36:354-63.	Pooled datasets of similar trials
George SZ, Zeppieri G, Jr., Cere AL, Cere MR, Borut MS, Hodges MJ <i>et al.</i> A randomized trial of behavioral physical therapy interventions for acute and sub-acute low back pain (NCT00373867). <i>Pain</i> 2008;140:145-57.	Included patients aged less than 18 years and also looked at effect modification over time
Haas M, Grouppe E, Muench J, Kraemer D, Brummel-Smith K, Sharma R <i>et al.</i> Chronic disease self-management program for low back pain in the elderly. <i>Journal of Manipulative & Physiological Therapeutics</i> 2005;28:228-37.	Intervention not delivered by therapist
Hagen EM, Svensen E, Eriksen HR. Predictors and modifiers of treatment effect influencing sick leave in subacute low back pain patients. <i>Spine</i> 2005;30:2717-23.	Outcome in subgroup analyses not a clinical measure of low back pain (return to work)

Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Independent evaluation of a clinical prediction rule for spinal manipulative therapy: a randomised controlled trial. <i>European Spine Journal</i> 2008;17:936-43.	Testing a clinical prediction rule
Jellema P, van der Windt DA, van der Horst HE, Twisk JW, Stalman WA, Bouter LM. Should treatment of (sub)acute low back pain be aimed at psychosocial prognostic factors? Cluster randomised clinical trial in general practice. <i>BMJ</i> 2005;331:84.	Look at effect modification over time
Jellema P, van der Roer N, van der Windt DA, van Tulder MW, van der Horst HE, Stalman WA <i>et al.</i> Low back pain in general practice: cost-effectiveness of a minimal psychosocial intervention versus usual care. <i>European Spine Journal</i> 2007;16:1812-21.	Outcome in subgroup analyses not a clinical measure of low back pain (cost-effectiveness)
Kool JP, Oesch PR, Bachmann S, Knuesel O, Dierkes JG, Russo M <i>et al.</i> Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. <i>Archives of Physical Medicine & Rehabilitation</i> 2005;86:857-64.	Outcome in subgroup analyses not a clinical measure of low back pain (days worked over 3 months)
Lamb SE, Lall R, Hansen Z, Castelnovo E, Withers EJ, Nichols V <i>et al.</i> A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. <i>Health Technology Assessment (Winchester, England)</i> /20;14:1-253.	HTA report. Secondary subgroups analyses paper published elsewhere and used instead (Underwood 2011)
Scheel IB, Hagen KB, Herrin J, Oxman AD. A randomized controlled trial of two strategies to implement active sick leave for patients with low back pain. <i>Spine</i> 2002;27:561-6.	Outcome in subgroup analyses not a clinical measure of low back pain (active sick leave)
Skargren EI, Carlsson PG, Oberg BE. One-year follow-up comparison of the cost and effectiveness of chiropractic and physiotherapy as primary management for back pain. Subgroup analysis, recurrence, and additional health care utilization. <i>Spine</i> 1998;23:1875-83.	Looked at an addition disorder (neck pain)
Skargren EI, Oberg BE, Carlsson PG, Gade M. Cost and effectiveness analysis of chiropractic and physiotherapy treatment for low back and neck pain. Six-month follow-up. <i>Spine</i> 1997;22:2167-77.	Looked at an addition disorder (neck pain)

Staal JB, Hlobil H, Koke AJ, Twisk JW, Smid T, van MW. Graded activity for workers with low back pain: who benefits most and how does it work? <i>Arthritis & Rheumatism</i> 2008;59:642-9.	Outcome in subgroup analyses not a clinical measure of low back pain (return to work)
Steenstra IA, Knol DL, Bongers PM, Anema JR, van MW, de Vet HC. What works best for whom? An exploratory, subgroup analysis in a randomized, controlled trial on the effectiveness of a workplace intervention in low back pain patients on return to work. <i>Spine</i> 2009;34:1243-9.	Outcome in subgroup analyses not a clinical measure of low back pain (return to work)
Thomas KJ, MacPherson H, Ratcliffe J, Thorpe L, Brazier J, Campbell M <i>et al.</i> Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain. <i>Health Technology Assessment (Winchester, England)</i> /1/10;9:iii-iv.	HTA report. Secondary subgroups analyses paper published elsewhere and used instead (Thomas 2006)
Toda Y. Impact of waist/hip ratio on the therapeutic efficacy of lumbosacral corsets for chronic muscular low back pain. <i>Journal of Orthopaedic Science</i> 2002;7:644-9.	Intervention not delivered by therapist (Corsets given to patients)
van Poppel MN, Koes BW, van der Ploeg T, Smid T, Bouter LM. Lumbar supports and education for the prevention of low back pain in industry: a randomized controlled trial. <i>JAMA</i> 1998;279:1789-94.	Intervention not delivered by therapist (Lumbar supports given to patients)

PART II

Statistical Methodology

Chapter 5

Review of subgroup methodology

5.1 Introduction

The previous chapter presented findings from a systematic review of the quality, conduct and reporting of subgroup analyses performed in randomized controlled trials (RCTs) of therapist delivered interventions for the management of non-specific low back pain (LBP). The review found that the majority of subgroup analyses performed in the area of non-specific LBP are only able to provide exploratory or insufficient findings. Moreover, the general content and reporting of subgroup analyses is rather poor. The review concluded that more novel statistical approaches need to be sought to better deal with the existing issues associated with the conventional approach.

Alternative methods for performing subgroup analyses or subgroup identification may exist elsewhere in other fields of research. The term subgroup analysis has different meanings for different people. Some regard it as the analyses of a defined subset or subpopulation of the entire data e.g. analysis of elderly individuals only. Some may consider it an approach to identify prognostic subgroups i.e. subgroups with high or

low outcome, whereas some understand it to be the identification of subgroups defined by baseline characteristics that moderate treatment effect. In this thesis, the latter is of interest however it is important to describe these other approaches to better understand the distinction. Though methods for identifying prognostic subgroups are not the primary goal, it is important to set the scene and also to review methods here that might be useful for finding moderators. This chapter will therefore present a review of statistical methods for conducting subgroup analyses or subgroup identification in general for a single trial based setting. Firstly the identification of prognostic subgroups will be described and discussed, followed by a description of methods for subgroup analyses to identify moderators of treatment effect. As highlighted in the previous chapter, one option to improve the power issue is to perform subgroup analyses on individual patient data (IPD) collected from several similar trials measuring similar outcomes. Therefore, this chapter will also briefly look at current methods used for performing subgroup analyses in IPD meta-analyses.

5.2 Methods to identify subgroups with high or low outcome

Regression modelling

In many fields of research, a 'subgroup analysis' refers to the identification of subgroups with high or low outcome i.e. prognostic subgroups. Such analyses are performed in different kinds of study designs such as RCTs, prospective cohort studies and retrospective studies. The identification of prognostic factors provides clinicians with valuable information to aid decision making and to help them better predict outcome. Depending on the outcome of interest associated with the field of research, an appropriate multivariate regression modelling approach is taken to identify and evaluate prognostic factors. For example, linear regression modelling is used when the outcome is continuous and normally distributed. There is no ideal approach to

identifying and modelling prognostic factors. However, two standard strategies are to either fit a full model that includes all potential prognostic factors or to use backward elimination (111). Inferences can then be made by observing the parameter estimates of the final fitted model as to whether any of the included factors are predictive of outcome. For example, imagine the final model was of the form

$$E(y) = 1.3 + 3.4 * Gender(Female)$$

where y is a continuous and normally distributed outcome, which here might, for example be a pain score on a scale from 0-100, where a higher score indicates worse pain. Also, suppose gender was found to be statistically significant in the model output, hence suggesting it is a prognostic factor. Then the regression coefficient for gender would suggest that females, on average, experience more pain than males i.e. the gender of an individual helps predict whether the response will be high or low.

Non-parametric methods

When the predictor is continuous, one may encounter situations where there is a non-linear association between the potential prognostic factor and the dependent variable i.e. the linearity assumption does not hold. In such a situation, one might consider using generalized additive models (GAMs) to model the non-linearity (112). Typically regression models such as linear and logistic regression work by modelling the X_i ($i = 1, \dots, m$) potential variables (including interaction variables) as a linear predictor of the form $E(Y) = \beta_0 + \sum \beta_i X_i$. GAMs work by replacing the linear predictor by an additive predictor of the form $E(Y) = \beta_0 + \sum f_i(X_i)$ with a monotonic link function to link $\beta_0 + \sum f_i(X_i)$ to the expectation of Y . Here, f_i ($i = 1, \dots, m$) are non-parametric smooth functions that are estimated from the data. This approach is relevant when a continuous covariate is of interest e.g. age. A general linear model can thus be

considered a special case of a GAM, where the link function is simply the identity function. The flexible nature of GAMs allows the assumption of linearity to be relaxed however extra caution must be taken not to over-fit the data (112).

Cluster analysis

An alternative approach is cluster analysis; an approach used in the field of data mining and machine learning for identifying subgroups or clusters of patients that are most similar in terms of outcome only, i.e. ignoring the predictors (113-115). Cluster analysis is based on mathematical formulation, whereby each individual is assigned to only one subgroup based on the Euclidean distance from an initial starting value or centroid. This method works by taking a large heterogeneous population and breaking it down into subgroups that maximize the between subgroup heterogeneity whilst minimizing the within subgroup heterogeneity. The number of subgroups to be formed has to be specified before running the analyses; however the choice of how many is a difficult task. Moreover, a starting point or centroid for each of the pre-determined subgroups also has to be specified. The following example illustrates how the cluster analysis algorithm works. Consider a sample of patients that have reported a pain score (0-100; higher score is worse). Assuming we pre-specify that we are interested in forming two subgroups. We therefore have to specify a starting point for each subgroup. Based on the pain score, let's give one of the subgroups a low starting point, say 10, and the other subgroup a high starting point, say 80. The method works by computing the distance of the outcome of each patient from the two subgroup starting points and then assigning each patient to the subgroup it is closest to. A natural measure of the distance for each patient in this example would be the squared error. Thus, in the first iteration, all patients will be assigned to one of two subgroups. The mean pain score in each of the subgroups is then computed and used as the starting point for the next iteration. The iterations continue until the starting points no longer

change. Once the subgroups have been formed, the outcome measure can be summarized within each subgroup and then labelled accordingly to best describe the subgroup e.g. good response and poor response. The final clusters can then be compared in terms of the characteristics of the patients within each cluster or subgroup. Any differences found in the characteristics of the two groups can aid in predicting outcome of new patients. For example, the cluster classified as having a good response may be on average younger than the cluster classified as having poor response. Thus, age would be a predictor of outcome for a new patient. A common and well documented problem associated with cluster analysis is that different solutions are often produced when different starting points or centroids are used. Thus there is no justification as to which solution is the correct or final solution.

Data mining methods

In the field of data mining, a number of data driven approaches exist that are an attractive alternative for performing subgroup analyses with the aim of identifying main effects i.e. subgroups that are predictors of outcome. Many of these are sophisticated data mining methods, such as support vector machines (SVM), neural networks, Bayesian networks and K-nearest neighbour classifiers, which specialize in discovering patterns and relationships between covariates and outcome within large datasets using algorithms. An initial concern with many of these complex methods is that there are numerous algorithms available for each of these methods. The majority of the data mining methods identified from the literature search do not look for treatment effect heterogeneity or interaction effects. Instead, they simply look for subgroups or subsets of the entire dataset with heterogeneous outcome. However, there are some data mining methods that do exist that aim to identify interactions. These methods will be described later on in section 5.3.2.

As mentioned earlier, the approaches in this section of the chapter aim to identify subgroups that differ in terms of a final outcome i.e. identifying prognostic subgroups. Though these methods are a form of subgroup analysis, they do not identify subgroups that have high or low treatment effects, which is the focus of this thesis. The next section will therefore describe several methods that can be used to identify and evaluate subgroups with high or low treatment effects i.e. differential subgroups effects.

5.3 Methods to identify subgroups with high or low treatment effects

5.3.1 Single factors

In general across many fields of research, it is of particular interest to perform subgroup analyses to investigate how the effect of one variable on an outcome variable of interest is moderated by the value of a third variable (moderator variable). In a clinical trial setting when comparing two treatments, the aim of subgroup analyses is to determine whether or not there is treatment effect heterogeneity in terms of an outcome between subgroups of patients. These subgroups are defined by the patient's individual baseline characteristics collected prior to randomization (56). Commonly, the baseline characteristics are investigated one at a time i.e. single factors. A number of statistical methods for evaluating single baseline factors will now be described.

Statistical test for interaction

Interaction tests as described in Chapter 3 are the most common approach for conducting subgroup analyses. There have been a number of different interaction tests proposed over the years for performing subgroup analyses, in particular, detecting treatment effect heterogeneity (116-118). However, in essence, subgroup analyses are performed by selecting an appropriate regression model depending on the outcome

variable and then subsequently adding an interaction term to the model. When the outcome is continuous and normally distributed then a linear regression model is used and when the outcome is binary then a logistic regression model is used. The baseline variable used to form the interaction term can either be categorical or continuous. When the baseline variable is continuous, it will require categorization that is clinically justified and sensible to allow for easier interpretation of the subgroup analyses results.

Qualitative interaction test

As mentioned in Chapter 3, interactions can be classified as being either quantitative or qualitative (43, 44). Just to recap, an interaction is said to be quantitative when the treatment effect is in the same direction (i.e. superior or inferior) within each subgroup but differs in terms of size. On the other hand, an interaction is said to be qualitative when the treatment effect in one subgroup is in the opposite direction compared to the other subgroup. Clinically, both types of interaction are of interest. Of the two, quantitative interactions are expected to occur whereas qualitative interactions are more difficult to find but are really important as they indicate which patients are harmed. Numerous qualitative interaction tests have been developed and proposed. Assuming there are two trial arms being investigated, these tests evaluate whether or not qualitative interactions exist across two or more disjoint subgroups. Gail and Simon initially proposed a likelihood ratio test (119) which was then compared in terms of power to a range test proposed by Piantadosi and Gail (120). The power comparison of both methods found that the likelihood ratio test had more power when the treatment was found to be harmful in several subgroups, whereas the range test had more power when the treatment was harmful in only a small number of subgroups (120). An extension of the range test was later developed by Li and Chan that outperformed the original range test and had greater power than the likelihood ratio

test when one treatment was found to be superior than the other in most subgroups (121). However, a limitation with qualitative interaction tests is that the disjoint subgroups must be pre-defined and formed prior to the analyses. For example, a continuous covariate such as age will have to be dichotomized or categorized prior to the analysis.

Non-parametric methods

When the distributional assumptions of a regression model do not hold or if there is some indication that they will be violated, then one may be inclined to utilize a non-parametric approach for performing subgroup analyses. Crump et al. developed and proposed two non-parametric tests for testing treatment effect heterogeneity. The null hypothesis for the first test is that the mean treatment effect for all subgroups defined by covariates is zero, whereas the null hypothesis for the second test is that the mean treatment effect conditional on the covariates is the same for all subgroups (122). Kraemer et al also proposed a non-parametric approach for performing subgroup analyses to test binary moderators by using area under the receiver operating curve (AUC) (123).

Graphical methods

Some statisticians argue that variables measured on a continuous scale should not be dichotomized as it reduces the power to detect any difference between the dichotomized variable and the outcome (124, 125). To avoid this issue and maintain as much power as possible, a number of graphical methods have been proposed that individually evaluate continuous variables. Royston and Sauerbrei introduced the multivariable fractional polynomials interaction (MFPI) method (126). The MFPI method involves using fractional polynomials to model the best fit of a continuous by binary interaction. A graphical representation of the continuous treatment effect

function would then be used to interpret any significant interactions found. Although the variable is kept continuous, ultimately a cut-point selection is still required having observed the graph to best define the subgroups. A similar method was also introduced by Bonnetti and Gelber called the subpopulation treatment effect pattern plot (STEPP) that investigates one or more subgroup variables of interest (127). It investigates interactions by using a graphical approach looking at overlapping subpopulations of patients and assessing the treatment effects across the subpopulations. However issues with this approach, as acknowledged by the authors, are multiplicity when considering other subgroups and also how to define groups when two or more overlapping subgroups show conflicting subgroups effects (127).

Bayesian approach

There are several concerns with the classical approach of hypothesis testing to perform subgroup analyses such as multiple testing and lack of power. Alternative Bayesian approaches have therefore been proposed that use shrinkage estimation techniques to perform subgroup analyses and identify differential subgroups. Despite the proposition of alternative Bayesian approaches, the classical approach to subgroup analyses and identification are still preferred and utilized. One of the reasons for this is that it can be quite challenging to specify an appropriate prior distribution when using a Bayesian approach. Bayesian methods for performing subgroup analysis are relatively new and are currently an ongoing area of research, including the extension of Bayesian subgroup analyses using shrinkage estimation in a meta-analysis framework (128).

5.3.2 Multiple factors

A particular limitation with the aforementioned differential subgroup analysis methods (section 5.3.1) is that they are typically employed to investigate patient characteristics

one at a time, whereas it is quite obvious that each patient has multiple characteristics. It is possible for some of the aforementioned methods e.g. regression based approaches, to consider multiple patient characteristics by investigating all possible treatment-covariate interactions in some sort of a stepwise manner. However, such an approach may not be sensible as it would drastically exacerbate the multiplicity issue due to the increased number of hypotheses being tested. For this reason, alternative approaches that incorporate multiple patient characteristics have been proposed in the literature. One approach is a risk-stratified analysis to create subgroups that can then be investigated for heterogeneity. This approach is described in more detail below. Furthermore, there are methods that exist in the field of data mining that can also be considered for performing subgroup analyses or subgroup identification. These methods do not make any distributional assumptions and as the name suggests, they mine the data with the aim of unearthing patterns or structures inherent within the data. These alternative candidate approaches, namely multivariable risk-stratified analysis and tree based methods, for subgroup analyses or subgroup identification that take into account multiple patient characteristics will now be described in more detail.

Multivariable risk-stratified analysis

To incorporate multiple patient characteristics when investigating treatment effect heterogeneity, Hayward et al suggested using a multivariable risk-stratified analysis. This method uses an adequate externally developed and pre-validated risk prediction tool that is readily available to form risk-stratified subgroups that are then tested for treatment effect heterogeneity (129). Although this approach is commendable, an issue with it however is that a pre-developed risk prediction tool is required but may not be available in the field of research where it is to be applied. Moreover, if a risk prediction tool is utilized and no treatment heterogeneity is found, then one possible reason for

this could be that the wrong subgroups are being investigated. This therefore might not be a suitable approach to pursue.

Tree based methods

Considering the multidisciplinary nature of the research areas in which subgroup analyses are applied, it is important that the methodology used is understandable, applicable and interpretable by clinicians and not overly complex. There is one such data mining approach in particular, tree based modelling, which is quite popular and commonly used in data mining as well as many other fields of research. Tree based methods are exploratory in nature and employ a rather intuitive technique referred to as recursive partitioning. Moreover, they do not require as many complex tuning parameters to be specified compared to many other data mining methods.

Furthermore, the basic principles behind this approach can be easily understood and the output easily interpreted.

Tree based methods allow us to explore the entire covariate space for relevant simple and complex interactions as opposed to conventional subgroup analyses which only explore covariates of interest one at a time. The method uses recursive partitioning, a non-parametric technique, that makes no assumptions about the functional form relating covariates to outcome. This therefore makes the method more robust to any violation of the assumptions made by the conventional regression approach to subgroup analyses (130). In brief, these methods rely on a splitting criterion to recursively form binary splits of the entire covariate space to identify homogenous subgroups that when compared are heterogeneous in terms of some outcome measure. A more detailed explanation as to how tree methods work will be provided in Chapter 6. The results are then displayed as a hierarchical tree structure that is easily interpretable to all. Classification and Regression Trees (CART) is probably the most

popular tree based method that is used today (131). Although the original tree based approach searches for heterogeneity in the outcome measure, there are now several advanced variants that have been recently proposed to identify subgroups of heterogeneous treatment effect (132-136). While this approach is completely exploratory in nature, it is a subgroup identification method that investigates subgroups defined by multiple characteristics that may have gone unnoticed using the conventional regression based approach.

5.4 Subgroup analysis methods for IPD from trials

Ordinary meta-analyses that synthesize aggregated data from several similar studies are a popular form of analyses. The methodology has been used for many years and hence is very well established. Individual patient data (IPD) meta-analyses on the other hand, regarded as the gold standard for meta-analyses, use the original individual patient data from each of the studies, which makes the analyses rather different. IPD meta-analyses have greater power and are particularly more useful compared to individual trials and ordinary meta-analyses methods when patient-level covariates are of interest rather than or in addition to just the mean effects. Though it is the ideal approach for performing meta-analyses, like any method, there are a number of challenges faced mainly to do with the approach being resource intensive (71). Performing IPD meta-analyses has only recently gained much popularity and therefore its methods are not as well established as ordinary meta-analyses methods.

When performing IPD meta-analyses, in particular subgroup analyses, the existing methodology requires that one takes either a two-step approach or a one-step approach. In the two-step approach, the subgroup analyses are carried out in each individual dataset separately to obtain subgroup effect estimates along with their respective variance estimates. These estimates are then synthesized in a similar

manner to when performing ordinary meta-analysis using existing techniques.

However, a problem with this approach is that often a simple fixed-effect pooling is used that does not incorporate the heterogeneity among the studies (137). This issue is overcome by using a one-step approach, a more flexible approach, which fits a single and simple hierarchical model with the inclusion of an interaction term (similar to the linear/logistic regression models) and also includes random effects to account for the between study heterogeneity.

Despite IPD meta-analyses having greater power when wanting to identify moderators of treatment effect, the current methodology used may not be ideal. The problem is that we only know how to extend the simpler models to the IPD meta-analyses setting. To be more precise, when either a one stage or a two stage approach is used for subgroup analyses, interaction tests are performed testing one patient characteristic at a time; they do not consider the multiple characteristics of patients. This is the exact same issue highlighted earlier with the methods used in single trials. Therefore, although IPD meta-analyses provide an ideal framework for subgroup analyses, there is a need for methodological development to incorporate multiple patient characteristics when performing subgroup analyses.

5.5 Discussion

The term subgroup analysis is interpreted differently by different people. The distinction between prognostic subgroups and differential subgroups was made at the start of this chapter where the latter is of interest in this thesis. Differential subgroup analyses are most commonly performed using a regression based approach with the inclusion of a treatment-covariate interaction to test for treatment effect moderation. This chapter performed a broad literature review to explore the wider literature for other proposed methods of performing subgroup analyses. The review process found a

number of methods that have been proposed to date for investigating subgroup effects and subgroup identification using individual patient characteristics and also using multiple characteristics as well. Moreover, IPD meta-analysis methods were briefly described since an IPD framework is ideal for performing subgroup analyses. However, it became apparent from this review that there is a need for methodological development in both the single trial case and IPD meta-analysis case to incorporate multiple patient characteristics when identifying subgroups because in general only the simpler regression type models with a single interaction effect are used.

Having reviewed the literature and identified various methods, it would not be possible to explore and evaluate every single method. Therefore, it is probably worthwhile at this stage to contrast what is written in current proposed guidelines for performing subgroup analyses to the findings from the systematic review of subgroup analyses in the area of low back pain presented in Chapter 4. One of the key recommendations in current proposed guidelines is that a clear distinction be made between pre-specified and post-hoc analyses where the former is for hypothesis testing (confirmatory analyses) and the latter for hypothesis generating (exploratory analyses). A limited number of pre-specified subgroups for investigation must be chosen using either a clear clinical justification or it must be based on findings from previous studies. As highlighted in chapter 3, if clinical justification is used then this could be quite subjective and it may be that important subgroups may go unnoticed. On the other hand if one were to base their choice of subgroups on findings from previous studies, then this may not be very wise if the quality, conduct and reporting of subgroup analyses in that particular field is poor; as was found in the systematic review in Chapter 4. For these reasons, it often makes sense to keep the subgroup analyses entirely exploratory in nature and use a method that investigates the entire covariate space such that no important subgroup effects go unnoticed. Any subgroups that are

identified from the exploratory analyses can then be tested in a future trial. Tree based methods, as described in this review, are one such method that can accommodate this. Furthermore, recent development of this methodology makes this a promising approach for subgroup identification in the context of clinical trials. The rest of this thesis will therefore focus on the evaluation, development and application of tree based methodology for identifying subgroups when using individual patient data from several similar trials.

A key constituent of tree based approaches is the utilization of a technique referred to as recursive partitioning. Therefore the following chapter will introduce the recursive partitioning methodology, followed by a description of the several advanced tree based method variants that have been proposed in the literature to date for performing differential subgroup analyses. A simulation study will be performed thereafter to assess the performance of these variant methods in detecting interactions in a single trial setting and the results presented.

Chapter 6

Introduction to Recursive Partitioning

6.1 Introduction

The previous chapter provided a broad review of the available methods for conducting subgroup analysis or subgroup identification in a single trial based setting as well as methods typically used for individual patient data (IPD) subgroup meta-analyses. Of the methods described in the review, it was identified that tree based methods are an attractive possibility for performing subgroup analyses using multiple patient characteristics. Whilst the application of tree based methodology is evident elsewhere as a valuable tool for identifying subgroups, its use in IPD meta-analyses and clinical trials of musculoskeletal disorders research is un-explored.

Tree based methods are a data driven approach from the field of data mining that use a simple intuitive technique called recursive partitioning. It is important to initially fully understand the underlying recursive partitioning methodology of the tree based methods before looking at advanced variants of this methodology and considering possible extensions to an IPD meta-analyses setting. This chapter will therefore start

by describing the recursive partitioning methodology along with its relevant steps. Since outcome measures in the area of low back pain research are mainly of the continuous type, the explanation of the recursive partitioning procedure in this chapter will be for the continuous outcome case (regression trees). It is important to note here that the recursive partitioning methodology described in this chapter will first focus on identifying subgroups of patients that differ in terms of outcome, thereafter recursive partitioning methods to find moderators of treatment effect will be considered. However in traditional subgroup analyses, the aim is to identify subgroups that differ in terms of treatment effect i.e. treatment effect heterogeneity. Therefore the sole purpose of section 6.2 is for the reader to gain a good understanding of the concepts and basic methodology of recursive partitioning for tree based methods. The advanced variants of this methodology that have been proposed in the literature to date to identify treatment effect heterogeneity will be explained thereafter in section 6.3.

6.2 Tree Based Methods

As highlighted in previous chapters, there are several limitations to the conventional approach to subgroup analyses. These issues include a lack of power, multiplicity, subgroup selection for analyses and consideration of multiple patient characteristics. Selection of subgroups to be investigated in conventional analyses is rather subjective. Also, the selected subgroups are investigated independently despite the fact that patients have multiple characteristics. Tree based methods resolve these issues by searching the entire covariate space to identify subgroups using multiple patient characteristics.

The earliest evidence of the application of recursive partitioning to create regression trees was in 1963 by Morgan and Sonquist at the University of Michigan, who proposed a method then referred to as the Automatic Interaction Detection method (AID)(138).

This method was primarily developed to analyse multiple covariates in terms of a single continuous response variable to detect interactions effects. However, this approach was very rarely implemented due to a serious problem, as highlighted by Breiman et al., of over-fitting and the results being unstable (131). In 1984, Breiman et al. then went on to extend the AID methodology to overcome these problems and proposed a method that is most commonly referred to today as the Classification And Regression Tree (CART) method (131). As the name suggests, the method is applicable to situations where the outcome is either categorical (classification trees) or continuous (regression trees). Since the proposal of CART, several other variants based on the CART methodology and application have been proposed. A limitation of the CART method is that there is no formal way of making statistical inference. Therefore some of the more recent advancements combine multiple regression analysis with regression trees to accommodate a formal way of making statistical inference. Some of these include methods for estimating regression models consisting of both main effects and interaction effects. These methods include multivariate adaptive regression splines (MARS) proposed by Friedman in 1991, the M5 algorithm proposed by Quinlan in 1992 and generalised unbiased interaction detection and estimation (GUIDE) proposed by Loh in 2002 (139-141). There are also methods that have been specifically developed or have the ability to detect treatment-effect heterogeneity. For example the Regression Trunk Approach (RTA) proposed in 2004 by Dusseldorp et al and the Simultaneous Threshold Interaction Modelling Algorithm (STIMA) proposed in 2010 by the same authors (132, 133). It is thus evident that recursive partitioning methodology has developed over the years and an excellent detailed review is provided by Zhang and Singer (2010) on modern recursive partitioning and its applications (130).

Initially, it is essential to understand the underlying methodology of recursive partitioning as it is a key component of tree based methods. This consists of three key steps:

- 1) Growing the initial fully grown tree
- 2) Pruning the fully grown initial tree to identify potential optimal sub-trees
- 3) Select the optimal sized sub-tree

These steps will now be described in the following sections.

6.2.1 Growing an Initial Fully Grown Tree

Tree based methods involve the use of a technique referred to as recursive partitioning; a non-parametric technique. Recursive partitioning is a process that, as the name suggests, recursively forms binary splits of the covariate space X_j , where j denotes the j^{th} covariate, with each new split forming two homogenous subgroups of individuals that when compared are most heterogeneous in terms of some response variable, say Y , where Y is either categorical or continuous (130).

The process initially starts at a unique root node, denoted by τ , that consists of the entire dataset i.e. all N individuals. A binary split of the root node is then created to form two new child nodes by choosing an optimal split point, say s , of a covariate X_j that is either continuous or categorical. How to find the optimal split point will be explained later on. If X_j is continuous then individuals with a value less than or equal to the optimal split point s are assigned to the left child node $\tau_L = \{i | i \in \tau, X_{i,j} \leq s\}$ and the remaining individuals with a value greater than the optimal split point s are assigned to the right child node $\tau_R = \{i | i \in \tau, X_{i,j} > s\}$, where $X_{i,j}$ is the value of the j^{th} covariate for the i^{th} individual. If X_j is categorical then the split point s will form two disjoint subsets, say A and B , such that individuals contained in subset A are assigned

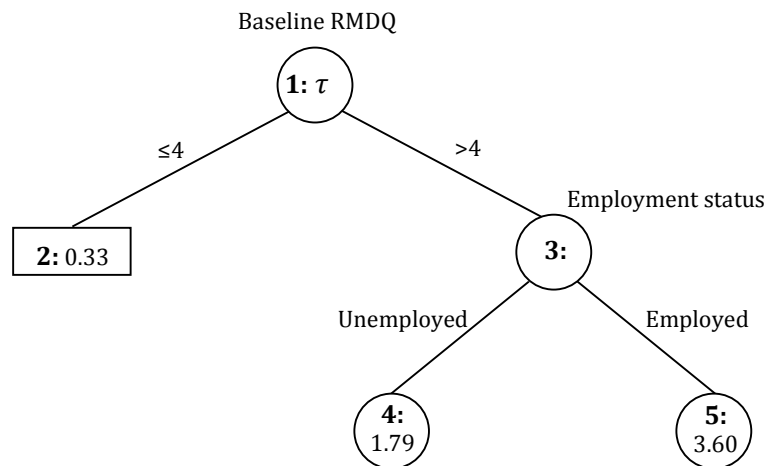
to the left child node $\tau_L = \{i | i \in \tau, X_{i,j} \in A\}$ and those contained in subset B are assigned to the right child node $\tau_R = \{i | i \in \tau, X_{i,j} \in B\}$. These newly formed child nodes are also referred to as internal nodes. The same binary splitting process at an optimal split point s of a covariate X_j is then applied to each of the newly formed internal nodes. This process of recursively creating binary splits of the internal nodes continues until either a formal stopping rule criterion is met or when no more splits can be made on the most recently created internal nodes, hence resulting in a fully grown initial tree. Choosing which split is best and when to stop growing a tree is described later on in section 6.2.3. Internal nodes that cannot be split any further are referred to as terminal nodes, where each terminal node is simply a homogenous subgroup in terms of the outcome and also a subset of the root node.

6.2.2 Tree growing example

A simple hypothetical example of the initial stages of a tree growing process with a continuous outcome (regression tree) has been illustrated in figure 6.1. This example assumes we have some data from a non-specific LBP RCT where the outcome is the change from baseline to 12 months in the Rolland and Morris Disability Questionnaire (RMDQ) and the aim is to identify predictors rather than moderators (28). The regression tree consists of three layers; the first layer contains just node 1 i.e. the root node τ , the second layer contains one terminal node (node 2) and one internal node (node 3) and finally the third layer contains two internal nodes (node 4 and node 5). In the first layer, the root node is split using the baseline RMDQ score at $s=4$ to form left (node 2) and right (node 3) child nodes, τ_L and τ_R respectively. The newly formed child nodes are then split in the same manner as the root node. Let's assume node 2 cannot be split any further due to some stopping criterion. Node 2 therefore becomes a terminal node; represented by a square box. This therefore suggests that the subgroup

of patients with baseline RMDQ \leq 4 would expect to see a mean change from baseline to 12 months of 0.33 in the RMDQ score. Node 3 can be split further using the binary covariate employment status (unemployed and employed) as it is the best split for this node out of all possible splits (optimal split). This split creates a new set of left and right child nodes (node 4 and node 5). What this means is that the subgroup of patients with baseline RMDQ $>$ 4 and who are also employed would expect to see a mean change from baseline to 12 months of around 3.60 in the RMDQ score. Whereas the subgroup of patients with baseline RMDQ $>$ 4 but who are also unemployed would expect to see a mean change from baseline to 12 months of 1.79 in the RMDQ score. Nodes 4 and 5 then become internal nodes, represented by circles, which then need to be searched to see if they can be split further. The binary splitting process continues in this way until all internal nodes cannot be split any further resulting in all the end nodes becoming terminal nodes.

Figure 6.1 – Example of a regression tree



6.2.3 Node splitting

Number of splits

Having explained the basic concepts behind growing an initial fully grown tree, there are two questions that need to be addressed:

- 1) How to choose which covariate out of all covariates to split and at what optimal split point s to split at?
- 2) How to decide when not to split an internal node (i.e. when to decide that an internal node is a terminal node)?

Before answering the first question, it is important to note here that the total number of potential split points varies at each stage of the tree growing process depending on whether the covariates are continuous or categorical. For a continuous covariate or discrete ordered covariate, the number of possible split points is simply one minus the total number of its distinct values. For example a continuous variable with 100 distinct values will have $100-1=99$ possible split points. If a covariate is categorical, with M categories, then there are $2^{M-1} - 1$ potential split points. For example, assume ethnicity is a categorical covariate with three categories; white, black and asian. Thus, it has three possible split points; white vs black and asian, white and black vs asian and finally white and asian vs black. The overall number of possible splits when initially splitting the root node is simply the summation of all possible splits from all covariates. Therefore to answer the first question, all potential splits for all variables are evaluated using a goodness-of-split criterion to decide which covariate to split and at what optimal split point s to split at.

Splitting criterion

The goodness-of-split criterion is typically an impurity function that measures the reduction in the heterogeneity of an outcome Y between two newly formed child nodes created when splitting an internal node. An impurity function is basically a function that quantifies how impure or heterogeneous two child nodes are having formed a split. In an ideal case, we would want an optimal split s to form two subgroups of individuals that are completely homogenous (pure) in terms of an outcome Y ,

however, we know this will be highly unlikely and that the child nodes will be ‘partially homogenous’ or ‘impure’. Moreover, the amount of node impurity will vary for all possible splits over all X_j covariates. Therefore an impurity function is evaluated for all possible splits to find the optimal split that maximises the reduction in impurity i.e. produces the most “pure” subgroups. The type of node impurity measure used depends on whether the response variable is continuous or categorical. In the case where the response is continuous, a natural option for a node impurity measure is the within-node sum of squares:

$$i(\tau) = \sum_{i \in \tau} (Y_i - \bar{Y}_\tau)^2 \quad (6.1)$$

where $i \in \tau$ are the individuals in node τ and \bar{Y}_τ is the mean of the response for those individuals in node τ . The goodness-of-split (impurity function) can therefore be calculated for a split s of an internal node τ to form left and right child nodes, τ_L and τ_R respectively, as follows:

$$\varphi(s, \tau) = \Delta i(s, \tau) = i(\tau) - i(\tau_L) - i(\tau_R) \quad (6.2)$$

Here the impurity function simply subtracts the impurity of the two child nodes from the impurity of its parent node. As mentioned earlier, the impurity function is evaluated over all possible splits to find the optimal split s for each covariate X_j . Subsequently the covariate that maximises $\varphi(s, \tau)$ i.e. the split that leads to the biggest difference between the means of the two groups, is chosen to form the new split. This procedure is recursively applied to the newly formed internal nodes at each stage to continue the tree growing process.

Stopping criteria

The second question is how to determine when to stop growing a tree. One possible solution, although not the best, is to implement some sort of stopping rule also referred

to as pre-pruning. For example you could set a minimum size for the number of individuals in a child node e.g. $n=10$ or 2% of the original sample size, such that it becomes a terminal node (i.e. stops splitting) if it goes below that number. However implementing such a stopping rule can be problematic resulting in the tree growing process either stopping too early (under-fitting) or too late (over-fitting) (131). Breiman suggested that no stopping rules be put in place and that an initial fully grown saturated tree T_0 is formed such that the nodes cannot be split any further i.e. all individuals in the node are identical. Such a tree is very well fitted to the available data but is rather unstable and relatively poor when predicting future data. Instead, simpler subtrees nested in T_0 may fit the data well enough but prove to be better predictors of future data hence making predictions more generalizable; however going through all possible subtrees could be a daunting task. Thus to overcome the problem of under or over fitting, improve model stability, improve the predictability of future data and to limit the choices of optimal subtrees, Breiman introduced the concept of post-pruning; a process analogous to backward stepwise regression that simply removes nodes that minimally contribute to the predictive accuracy of the tree (130). The next section will describe the post-pruning process which will be referred to as just “pruning” from now onward.

6.2.4 Pruning

The pruning process initially starts with a fully grown tree. The procedure then iteratively removes branches that least contribute to the predictive accuracy of the tree to form a sequence of potentially optimal nested subtrees from which the best optimal subtree T^* is selected. The best optimal subtree is the subtree that minimises the overall predictive error; explained in more detail in 6.2.5. This concept in the area of

tree based methods is referred to as cost-complexity pruning. For a fully grown tree, T , the total tree cost-complexity is defined as:

$$R_\alpha(T) = R(T) + \alpha|\tilde{\tau}| \quad (6.3)$$

where $R(T)$ is a measure of the quality of a tree or the tree cost, $\alpha(\geq 0)$ is the complexity parameter (the cost of an additional single terminal node), and $|\tilde{\tau}|$ is the number of terminal nodes in the fully grown tree. Here the cost of the total tree $R(T)$, measured by the quality of its terminal nodes, is penalized further with respect to the complexity of the tree where the complexity is measured simply by the size of the tree i.e. the number of terminal nodes $|\tilde{\tau}|$. The complexity parameter α is a positive continuous real number where each value of α may lead to a different subtree that minimizes the cost-complexity. Breiman et al were able to show that every value of α has a unique subtree of the fully grown tree that minimizes the cost-complexity, thus there are a finite number of subtrees corresponding to a infinite number of complexity parameter values (142). Therefore, instead of searching through every possible subtree for each value of α to find the subtree with minimal cost-complexity using (6.3), Breiman and colleagues proposed an algorithm that created a sequence of complexity parameter values. This algorithm utilises a function $\alpha(\tau)$ to estimate the complexity parameter

$$\alpha(\tau) = \frac{R^s(\tau) - R^s(\tilde{\tau}_\tau)}{|\tilde{\tau}_\tau| - 1} \quad (6.4)$$

where $\tilde{\tau}$ is the set of all terminal nodes and $\tilde{\tau}_\tau$ is the set of offspring terminal nodes of the internal node τ , $R^s(\tau)$ is the resubstitution cost of the internal node τ and $R^s(\tilde{\tau}_\tau)$ is the resubstitution cost of the offspring terminal nodes $\tilde{\tau}_\tau$. Breiman et al. used the resubstitution cost to help prune back a tree. It is called the resubstitution cost because the same data used to build the tree are again used to estimate the cost of a tree. The function numerator $R^s(\tau) - R^s(\tilde{\tau}_\tau)$ in (6.4) basically compares the cost of a node τ to

the total cost of the terminal nodes in the branch connected to τ , denoted by $\tilde{\tau}$. Since we are describing regression trees here, i.e. the response is continuous; the pruning algorithm uses the sum of squared errors (SSE) as a measure of the resubstitution cost to prune back a tree ($R^s(\tau) = \sum_{i \in \tau} (Y_i - \bar{Y}_\tau)^2$). Note that the SSE was also used as the impurity function in the tree growing process described in the previous section.

Now that the components of the complexity parameter estimating function $\alpha(\tau)$ have been described, the steps of the pruning algorithm for determining the first subtree can now be explained. The algorithm consists of the following steps:

- 1) Let T_0 be a fully grown tree. Compute the estimate of the complexity parameter α using the function $\alpha(\tau)$ (see (6.4)) for all internal nodes (i.e. $\forall \tau \notin \tilde{\tau}$) of the initial fully grown tree T_0
- 2) Find the internal node with the smallest value of $\alpha(\tau)$ and remove (prune) all subsequent branches connected to this node. This internal node therefore becomes a terminal node. The resulting tree thus forms the first subtree T_1 corresponding to the complexity parameter estimate α_1 , as estimated by $\alpha(\tau)$.
- 3) Repeat steps 1) and 2) using T_1 as the initial tree.

Steps 1) and 2) from the above algorithm are continuously repeated using the previously formed subtree as the initial tree of the next iteration. The value $\alpha(\tau)$ computed for each internal node (step 1) reflects how much additional predictive accuracy the branch connected to node τ contributes to the tree. Hence, larger values of $\alpha(\tau)$ indicate greater contribution. Therefore, each iteration of the pruning procedure removes the branch that least contributes to the trees predictive accuracy thus forming an increasing sequence of complexity parameter estimates $\alpha_0 < \alpha_1 < \alpha_2 < \dots < \alpha_m$,

where $\alpha_0 = 0$ for the fully grown tree i.e. there is no additional cost for extra terminal nodes hence a fully grown tree is the best predictor. Furthermore, the complexity parameter sequence α_m ($m = 0, 1, \dots, m$) corresponds to a sequence of nested optimal subtrees $T_0 \supset T_1 \supset T_2 \supset \dots \supset T_m$, where each subsequent subtree in the sequence is a subtree of the previous tree i.e. $T_{m-1} \supset T_m$. The algorithm continues until the final subtree T_m in the sequence is just the root node. The optimal subtree T^* is then selected from the sequence of nested optimal subtrees. How to select the optimal subtree is described in the next section. This pruning procedure is quite often referred to as weakest-link pruning.

6.2.5 Selecting the optimal sub-tree

The previous sections of this chapter have described in detail the tree growing process. It was shown how to initially form a fully grown saturated tree that over-fits the data followed by a pruning procedure to remove unwanted branches that minimally contribute towards the predictive accuracy of the tree, thus forming a sequence of nested optimal subtrees of which one final optimal subtree T^* is to be selected. In order to select an optimal subtree, the predictive accuracy of each of the nested subtrees needs to be evaluated. This is done by initially computing the resubstitution SSE cost (continuous response) for each subtree and then selecting the optimal subtree that has the smallest cost. When no test data are available the original data used to generate the tree can be used to evaluate the predictive accuracy. This is referred to as the resubstitution estimate as the same data are being re-substituted. However, an identified problem with the resubstitution SSE cost estimates for the subtrees is that they tend to be biased downwards. This bias occurs because the same data are used to estimate the SSE cost hence the cost is lower due to better fitting subtrees, thus resulting in the final chosen optimal subtree T^* being larger than optimal (130). To

overcome this problem, the predictive accuracy of each of the subtrees needs to be estimated by applying it to an independent test dataset, if it is available. The individuals in the test dataset are sent down each of the subtrees and the cost of each subtree estimated from which the optimal subtree with minimal cost is selected.

Usually an independent test set is not available and so to overcome this, an intuitive resampling method can be used, namely, V -fold cross-validation. This method works by firstly separating the entire dataset into V equal pieces; typically V is set to 5, 10 or 25. If the dataset was split into 5 pieces (i.e. $V=5$) for example, then it would be referred to as 5-fold cross-validation. The next step is to hold out one of the pieces, say piece v , $v=1, \dots, V$, to form a test sample and use the remaining pieces as a training sample. So first of all, $v=1$ will be held out as the test sample and $v=2, \dots, V$ will be combined together to form the training sample. A fully grown tree is then formed using the training sample and then subsequently pruned using each of the complexity parameters from the sequence obtained when pruning the original fully grown tree using the entire dataset. This forms a sequence of nested subtrees that are then each applied to the test sample ($v=1$) to form unbiased estimates of the SSE cost, thus completing the first iteration. Similarly, the next iteration works by holding out $v=2$ as the test sample and using $v=1, 3, \dots, V$ as the training sample. In the same way, the iteration continues until each individual piece $v=1, \dots, V$ has been held out so that ultimately there are V sequences of nested subtrees with their corresponding SSE costs. Finally, an estimate of the SSE cost is then calculated by averaging the cost over all V folds for each complexity parameter α_m in the sequence. Now that these costs have been estimated, all that remains is to select the final optimal subtree T^* .

There are two ways in which this can be done. The first way is to choose the complexity parameter corresponding to the smallest average SSE cost as the optimal complexity

parameter value α^* . Therefore the subtree from the sequence of optimal subtrees that corresponds to this particular optimal complexity parameter value α^* is chosen as the optimal tree T^* . However, the final optimal subtree selected in this way can still be unstable. The second approach avoids this issue by using a method known as the 1-SE rule; an alternative method proposed by Breiman et al. that takes into consideration the standard errors of the SSE costs estimated from cross-validation. Not only does this reduce instability but it also selects the simplest subtree whose predictive accuracy is similar (within 1 standard error) to that of the tree selected from the cross-validation procedure. Let $R^{cv}(T_m)$ denote the cross-validation estimates of the SSE cost given by

$$R^{cv}(T_m) = \frac{1}{V} \sum_{v=1}^V R^s(T_{m,v}) \quad (6.5)$$

where T_m is the m^{th} subtree corresponding to the complexity parameter estimate α_m , $v = 1, \dots, V$ denotes the index of the fold used as the test set in the cross-validation and thus $R^s(T_{m,v})$ is the SSE cost of the m^{th} subtree obtained using the v^{th} fold from the cross-validation corresponding to the complexity parameter estimate α_m . Then the variance for the SSE cost estimate of the m^{th} tree can be computed as follows:

$$\text{Var}(R^{cv}(T_m)) = \frac{1}{N} \sum_{i=1}^N \left[(Y_i - \hat{Y}_{i,v})^2 - R^{cv}(T_m) \right]^2 \quad (6.6)$$

where Y_i and $\hat{Y}_{i,v}$ are the observed and predicted value for the i^{th} individual in the v^{th} test set respectively. The predicted value $\hat{Y}_{i,v}$ for an individual is simply the mean outcome in the terminal node they are assigned to having sent them down the tree. Having computed the variance, we can go on to compute the standard error:

$$SE(R^{cv}(T_m)) = \sqrt{\frac{\text{Var}(R^{cv}(T_m))}{N}} \quad (6.7)$$

It is important to note here that this is a heuristic approximation of the standard error estimate since it incorrectly assumes that the predicted values for all i individuals are independent. Having computed the SE, all that remains is to select the optimal tree using the 1-SE rule. This is done by firstly summing the smallest SSE cost estimate (computed using V-fold cross-validation) with the smallest SE estimate. The smallest subtree whose SSE cost estimate is no larger than this summation is chosen as the final optimal tree T^* . The selection of T^* using the 1-SE rule has been empirically shown in most cases to be of better quality compared to choosing the tree with minimum cross-validated SSE cost (referred to as the 0-SE rule) (130).

6.2.6 Interpretation of the optimal tree T^*

Once a final optimal tree T^* has been determined, we can easily make inferences about any subgroups inherent within the data by simply summarising the terminal nodes. When the response is continuous, subgroups identified using a regression tree are summarised by simply averaging the response for the individuals within each of the terminal nodes, thus creating a piecewise constant prediction function.

6.3 Advancements of recursive partitioning methodology to detect moderators of treatment effect

Since the formulation of the CART approach as proposed by Breiman, there have been several advanced variants of the CART type routine proposed in the literature. The majority of these variants focus on finding marginal effects rather than differential treatment effects; where the latter is of interest in this thesis. There are a handful of methods that are either designed specifically, or are capable of detecting differential subgroup effects. In particular, these methods are the Interaction Tree (IT), Regression Trunk Approach (RTA), Simultaneous Threshold Interaction Modelling Algorithm

(STIMA) and Subgroup Identification based on a Differential Effect Search (SIDES) (132, 133, 135, 136).

The IT method was proposed by Su et al to specifically detect treatment-covariate interactions i.e. moderators of treatment (136). The IT approach is similar to the CART approach but uses a different splitting criterion that specifically looks for maximal interaction effects. The RTA and STIMA approaches were both proposed by Dusseldorp et al with the aim of estimating a linear regression model as well as searching for interactions. The RTA approach, proposed in 2004, uses a three phase procedure for specifically detecting treatment-covariate interactions in addition to estimating a linear regression model. The authors then proposed the STIMA approach in 2010 as an improved version of the RTA method. The STIMA method is a more generalised method compared to RTA as it detects all interactions and not just treatment-covariate interactions. Moreover, it simultaneously estimates a linear regression model and a tree model whereas the RTA method uses three phases to do this. For STIMA to detect treatment-covariate interactions only, it requires the user to force the first split on treatment when growing the tree. As STIMA is an improved version of RTA, only the STIMA method will be considered further in this thesis.

The objective of the IT and STIMA procedures is to identify subgroups that most differ in terms of treatment effectiveness; hence aiming to find subgroups with a large treatment-covariate interaction. Moreover, both of these approaches follow a similar procedure to CART in that they initially grow a tree, prune the tree and finally select the optimal tree from the set of pruned subtrees. The trees produced by IT and STIMA tell us how many and what types of interactions are inherent within the entire dataset. Moreover, all of the data are sent down the tree and are contained within the terminal nodes. Therefore, inferences can be made about the entire dataset by observing the

terminal nodes. The terminal nodes can tell us which subgroups benefit the most, which subgroups benefit less and also which subgroups, if any, are harmed. The SIDES procedure on the other hand is rather different to IT and STIMA. At each level of the tree, SIDES also searches for subgroups with the largest treatment-covariate interaction. However, each time an interaction is found, the subgroup with the larger treatment effect is retained and the remainder of the data disregarded. Hence, the aim of SIDES is to identify subgroups in which the treatment arm outperforms the comparator arm. Furthermore, the SIDES procedure does not follow the typical CART type procedure, as will be explained later on, but does use recursive partitioning methodology. The difference between SIDES compared to IT and STIMA becomes clearer when considering the final tree obtained. Each of the terminal nodes of the SIDES tree defines a different candidate subgroup with enhanced treatment effect. It is quite likely that individuals in one candidate subgroup may also be contained in another candidate subgroup. Thus these subgroups are subsamples (possibly overlapping subsamples) of the entire dataset therefore inferences can only be made about subsets of the entire data. Moreover, unlike the IT and STIMA trees, the subgroups identified by SIDES can only conclude benefit. Although the aim of SIDES is different to the aims of IT and STIMA, from a clinical perspective, both of these aims are quite important. These methods will now be described in more detail.

6.3.1 IT method

Splitting criterion

The IT method works in a similar way to the CART procedure but instead uses a splitting criterion that detects treatment-covariate interaction. Suppose we split a node τ at some split point s to form two child nodes; τ_L and τ_R respectively. Also assume that

we have a continuous response Y and a treatment indicator, say Trt (control=0 and intervention=1). We can display this split as follows

	τ_L	τ_R
$Trt=0$	$\bar{Y}_0^L \ s_1^2 \ n_1$	$\bar{Y}_0^R \ s_2^2 \ n_2$
$Trt=1$	$\bar{Y}_1^L \ s_3^2 \ n_3$	$\bar{Y}_1^R \ s_4^2 \ n_4$

The terms in the first quadrant above $(\bar{Y}_0^L \ s_1^2 \ n_1)$ represent the sample mean, sample variance and sample size respectively in the left child node τ_L for when $Trt=0$. The other quadrants are interpreted in the same manner. Now that two nodes have been formed, all that is required is to determine the treatment effect heterogeneity between both nodes. This requires that we compare the treatment effect in the left node $(\bar{Y}_1^L - \bar{Y}_0^L)$ with the effect in the right node $(\bar{Y}_1^R - \bar{Y}_0^R)$. In other words, we are interested in the treatment-covariate interaction. The authors therefore proposed a splitting criterion for evaluating the interaction effect for each potential split in the tree growing process. The proposed splitting criterion is given by

$$G(s) = \left(\frac{(\bar{Y}_1^L - \bar{Y}_0^L) - (\bar{Y}_1^R - \bar{Y}_0^R)}{\hat{\sigma} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}} \right)^2 \quad (6.8)$$

where $\hat{\sigma}^2 = \sum_{i=1}^4 w_i s_i^2$ is a pooled estimator of the constant variance and where $w_i = \frac{(n_i-1)}{\sum_{j=1}^4 (n_j-1)}$ (136). Thus when growing the tree, the splitting criterion $G(s)$ is evaluated for every single potential split and the split with the maximum $G(s)$ is chosen as the best split. In other words, the split that gives the largest interaction effect is chosen as the best split.

Growing a tree – parameter specifications

A fully grown tree, say T_0 , is grown using the aforementioned split function until either some stopping criterion is met or the tree cannot split nodes any further. The code for the IT approach was very kindly provided by the author Dr Xiaogang Su. When applying the IT procedure, it requires a couple of user-defined parameters to be specified to aid the tree growing process. These include the minimum node size i.e. the minimum number of individuals in any given node, and the maximum depth of a tree i.e. how many levels the tree has (the complexity of a tree).

Pruning

The IT method uses weakest link pruning, similar to CART (see equation 6.4), to determine the complexity parameter values and the associated subtrees of the fully grown tree T_0 . Thus the function

$$\alpha(\tau) = \frac{\sum_{\tau \in \tau - \tilde{\tau}} G(\tau)}{|\tau - \tilde{\tau}|}$$

is evaluated for every internal node τ where the numerator $\sum_{\tau \in \tau - \tilde{\tau}} G(\tau)$ is the overall amount of interaction of the internal nodes in the branch connected to the internal node τ and the denominator $|\tau - \tilde{\tau}|$ is the number of internal nodes in the branch connected to τ . The internal node with the smallest value of $\alpha(\tau)$ is pruned i.e. the branch connected to the internal node is removed and the internal node itself becomes a terminal node to form the first subtree T_1 . In the same way as CART, this process continues to form a sequence of subtrees until we are left with just the root node.

Selecting the best tree

The quality or performance of an interaction tree, say T , is assessed using an interaction complexity measure; similar to the CART procedure (see equation (6.3)).

The interaction- complexity measure is given by

$$G_{\alpha}(T) = G(T) - \alpha \cdot |T - \tilde{T}|$$

where $G(T) = \sum_{\tau \in T - \tilde{T}} G(\tau)$ is the sum of the interaction in the internal nodes of the tree T and $|T - \tilde{T}|$ is the number of internal nodes in tree T . The above interaction-complexity measure can be evaluated for every subtree and the one with the maximum value is chosen as the best tree. What remains is the selection of the complexity parameter value α , i.e. the penalty for additional splits. This can be done using V-fold cross-validation as described in section 6.2.5, using the 1-SE rule to determine the best tree size and thus the complexity parameter value. The authors also recommend a bootstrapping method, used by LeBlanc et al, as an alternative option for validating the trees (136, 143).

Interpretation of the final tree

Once a final tree has been selected, the interpretation is rather straightforward. The outcome of interest, which in this case is the treatment effect, is computed for each of the terminal nodes of the final selected tree. The conclusions are then based on the comparison of the terminal nodes summaries. For example, say that a single one-way interaction effect exists in a dataset, then the final chosen interaction tree should consist of a single split of the root node i.e. two terminal nodes. Thus, the difference in the treatment effect between the two terminal nodes should be equivalent to the size of the interaction effect.

6.3.2 STIMA method

Splitting criterion

The STIMA method simultaneously estimates a linear regression model whilst searching for interaction effects by growing a tree. As mentioned before, the STIMA

method is more generalised compared to the previously proposed RTA method because it searches for all types of interaction. Therefore, in order for the method to search specifically for treatment-covariate interactions, the user has to force the first split on the treatment variable. It may be difficult to see how forcing the first split allows the tree to search for treatment-covariate interactions; however this will be explained later on.

The first step in the algorithm requires a linear regression model to be fitted using all covariates in the root node. Thereafter, all possible split points, say s , for all covariates are searched to identify the split that generates the largest effect size when comparing the model that was fit before split s to the model after adding split s as an indicator variable. The effect size is determined using

$$f_s^2 = \frac{(\rho_s^2 - \rho_{s-1}^2)}{(1 - \rho_s^2)} \quad (6.9)$$

where $\rho_{s-1}^2 = \frac{\sum_i (\hat{Y}_{i(s-1)} - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$ is the squared multiple correlation coefficient of the model before split s i.e. $s-1$, thus $\rho_s^2 = \frac{\sum_i (\hat{Y}_{is} - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$ is the squared multiple correlation coefficient of the model having split the node on s (133). This effect size essentially measures the increase in the variance-accounted-for (VAF) from the model without split s i.e. $s-1$, to the model with split s . The split point s that induces the largest relative increase in VAF is chosen as the next best split.

Growing a tree – parameter specifications

The aforementioned splitting criterion for STIMA is used to grow a full tree, which the authors also refer to as a regression trunk. Each iteration of the tree growing process evaluates the effect size for every single split of every single variable within every single node choosing only a single best split at the end of the iteration. This is different

to the IT approach which searches each node separately and creates a split for each node if it is permissible, whereas STIMA searches all nodes and forms just the one split. If a best split is found, it is added to the linear regression model as an indicator variable and then the model is re-estimated. This process is continuously repeated until either a stopping criterion is met or if the tree cannot find any more effective splits.

The code for the STIMA approach is readily available on the author's webpage (Elise Dusseldorp - <http://www.elisedusseldorp.nl/>). When applying the STIMA procedure, it requires a number of user-defined parameters to be specified to aid the tree growing process. These include the minimum node size, the maximum number of splits, the column index of the variable in the dataset on which to force the first split of the tree (if required) and finally the specification of the value V for the V-fold cross-validation procedure. The authors recommend using either 5-fold or 10-fold for the cross-validation procedure.

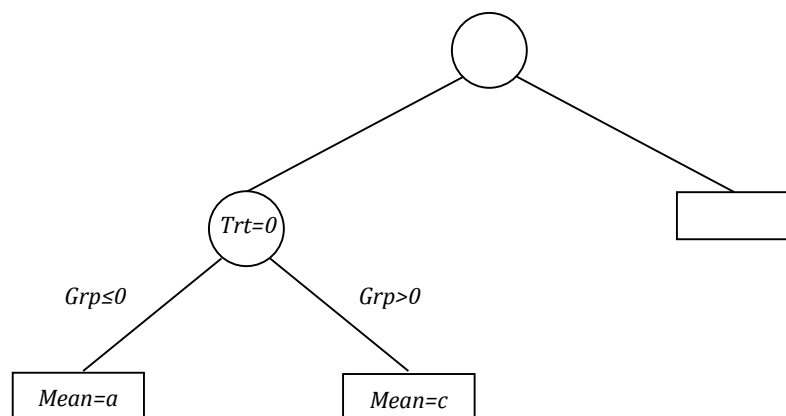
Forced first split on treatment

In order to detect treatment-covariate interactions, the first split must be forced on the treatment variable. To better understand this, consider an example of a single one-way interaction between a binary treatment variable, say *Trt*, and a binary baseline covariate, say *Grp*, with a *Trt*Grp* interaction effect. We can display the interaction by

	<i>Trt=0</i>	<i>Trt=1</i>
<i>Grp=0</i>	<i>a</i>	<i>b</i>
<i>Grp=1</i>	<i>c</i>	<i>d</i>

where a , b , c and d are the cell means. The interaction effect using the above is computed by $(a-b)-(c-d) = a-b-c+d$. After forcing the first split on treatment, represented by the bold dashed line, the method then goes onto search all splits for every covariate within all nodes to identify the next best single split. In this example, it can only split on Grp within both treatments. If the $Trt=0$ node is further split by Grp , then the means in the two newly formed child nodes will be a and c respectively. Hence the difference in means between the two child nodes is $a-c$. Similarly, if the $Trt=1$ node is further split by Grp , then the means in the two newly formed child nodes will be b and d respectively. Hence the difference in means between the two child nodes is $b-d$. Since the STIMA method makes just a single split per iteration, of the two splits, the method will select the split with the largest difference in cell means and then add it to the linear regression model as an indicator variable and the model re-estimated. If the next best split in this example is in the $Trt=0$ node, then one might assume that the same split is mirrored in the $Trt=1$ node in the next iteration, however this is not required since the model has been updated and re-estimated to incorporate the detected $Trt*Grp$ interaction effect; thus only one split is made. Figure 6.2 better illustrates the STIMA tree for the above example of a single $Trt*Grp$ interaction where the best split is made using node $Trt=0$.

Figure 6.2 – Example of STIMA tree for one-way interaction



Selecting the best tree

Once a full tree has been grown, STIMA then applies V-fold cross validation to select the best tree size i.e. the best choice of s ; the number of splits in the tree. Although CART and IT methods require a pruning procedure i.e. weakest link pruning, to obtain a sequence of subtrees, such a procedure is not required by STIMA. The reason for this is because STIMA is essentially a set of nested models that are defined by the number of splits s it has. Therefore, these nested models can be thought of as the sequence of subtrees obtained by the CART and IT pruning procedures. For example, a tree with four splits means that there are five candidate nested models to choose from; four models with interaction terms and the null model. Thus, the selection of the best size model determined by the number of splits s is done by using V-fold cross-validation in the same way as CART. To briefly explain, if a fully grown tree on the full dataset has three splits ($s=3$), then the cross-validation procedure will specify that the maximum number of splits formed when growing a tree on the training data is also three. This will result in four models being fitted, including the null model. The test sample is then used to obtain the predictive error of each model. The process is repeated V times in order to complete the V-fold cross-validation. In order to obtain stable estimates for the averaged predictive error for each tree from the cross-validation, the authors suggest repeating the cross-validation procedure several times and then averaging the results from which the best tree size is determined. For CART and IT, the best tree size is determined using the 1-SE rule, as described in 6.2.5. However the authors of STIMA suggest using a rule that depends on the sample size to improve the performance of STIMA. They suggest a 0.80-SE rule be used for sample sizes of less than 300 and a 0.50-SE rule used otherwise.

Interpretation of the final tree

Once the STIMA procedure has selected a final tree from the cross-validation procedure, the final tree can then be interpreted. This can be done either by observing the model output to see what the estimated interaction effects are for the interactions detected by the procedure. Furthermore, a plot of the STIMA tree (or regression trunk) can be observed to ease the understanding and interpretation of the subgroups identified.

6.3.3 SIDES method

Splitting criterion

The goal of the SIDES approach is different to that of IT and STIMA in that it aims to detect subgroups of patients with enhanced treatment effect. The IT and STIMA methods grow trees that represent interactions or differential treatment effects that have been detected. The SIDES method also searches for large differential treatment effects however once an optimal split is identified, it retains the subgroup with the enhanced treatment effect and disregards the rest of the sample. The SIDES procedure evaluates all possible splits for every covariate using a splitting criterion that computes the differential effect between the two subgroups formed by a split. The splitting criterion computes a p-value for all searched splits and is of the form

$$p = 2 \cdot \left[1 - \Phi \left(\frac{|Z_{E1} - Z_{E2}|}{\sqrt{2}} \right) \right], \quad (6.10)$$

where Z_{E1} and Z_{E2} are the one-sided hypothesis test statistics computed for the two subgroups respectively and $\Phi \left(\frac{|Z_{E1} - Z_{E2}|}{\sqrt{2}} \right)$ is the cumulative distribution function of the standard normal distribution (135). The one sided test statistics are obtained from a simple linear regression model with just the treatment variable as the covariate where

the test statistics are computed by dividing the treatment effect estimate by its standard error (t-statistic).

Multiplicity adjustment

Like all recursive partitioning methods, the SIDES algorithm requires all splits to be searched for every covariate at each iteration. A well-known issue with this is that the procedure has a greater probability of selecting a covariate with a larger number of levels. This issue is referred to as variable selection bias (144, 145). Thus to adjust for this, the authors introduced a Sidak-based multiplicity adjustment given by $1 - (1 - p_i)^{G^*}$, where p_i is the unadjusted p-value obtained using the splitting criterion (equation 6.10) for the i -th split and G^* is the effective number of splits (135). The effective number of splits basically incorporates the correlation between the p-values for any single covariate and is estimated by $G^* = G^{1-\bar{r}}$, where G is the number of splits for a given covariate and \bar{r} is the mean correlation of the p-values computed using the splitting criterion having evaluated all splits. Thus for a given covariate, the associated p-values for all potential splits are adjusted to reduce the variable selection bias.

Complexity control

The complexity of a tree simply refers to the size of a tree usually defined by the number of terminal nodes it has. Both the IT and STIMA approaches use pruning and cross-validation to control for the complexity of a tree and thus select the best tree size respectively. The SIDES approach on the other hand controls the complexity of a tree during the tree growing process by using a relative improvement parameter at each level of the tree. The relative improvement parameter determines whether or not a child node with a large positive treatment effect should become a parent node for the next iteration of the algorithm. A child node only becomes a parent node if there is

some improvement in the p-value of the child node compared to the p-value of the parent node from which it came from. The split is made if

$$p_c \leq \gamma \cdot p_p, \quad (6.11)$$

where p_c is the p-value of the child node, p_p is the p-value of the parent node and γ is the relative improvement parameter which ranges from 0 to 1. A small value of γ makes the procedure much more selective; only selecting subgroups with a very small p-value. Conversely, a value of 1 for γ is less restrictive and thus allows the procedure to search a wider covariate space i.e. build a bigger tree.

Each level of the SIDES tree has an associated relative improvement parameter. For example, if some restriction was only required for the first three levels of the tree, then one would need to specify γ_i for stage i . The relative improvement parameters can be user specified. If none are specified, they can be determined using 5-fold cross-validation to try and find the optimal combination. The cross-validation procedure is implemented in the same way as described in section 6.2.5. This requires the user to define the grid space that needs to be searched for the optimal combination of the γ parameters. For example, say we want to define a subgroup using three covariates only i.e. a tree with three levels, then we need to specify the values for γ_1, γ_2 and γ_3 to define the grid that needs to be searched. Let's say we specify that γ_1 goes from 0.1 to 1.0 in increments of 0.1 and both γ_2 and γ_3 go from 0 to 1 in increments of 0.1, then the grid is formed using all combinations of γ_1, γ_2 and γ_3 . Thus in this example, the first combination would be $\gamma = (0.1, 0, 0)$ and the last combination would be $\gamma = (1, 1, 1)$. Therefore, the 5-fold cross-validation works by applying the SIDES procedure to the training sample for each combination of γ in the grid. The best subgroup with the smallest treatment effect p-value is then identified and the same subgroup evaluated in the test sample and the corresponding p-value recorded. This is then repeated five

times for each fold in the cross-validation. Finally, the p-values are averaged across the 5-folds for each combination in the grid and the combination with the smallest average p-value is chosen as the optimal combination to be used by SIDES. Though this approach is good for finding the optimal combination, it can be rather time consuming if there are a large number of levels for which a relative improvement parameter is required. Therefore, the authors recommend only searching for the optimal combination for the first three levels of the tree.

Growing a tree – parameter specifications

Now that the splitting criterion, multiplicity adjustment and complexity control have been defined, the algorithm for growing a SIDES tree will now be described. The authors very kindly provided me with the coding for the SIDES method written in R software. When applying the SIDES procedure, it is required that a number of parameters be specified to aid the tree growing process. Like the IT and STIMA methods, SIDES also requires that the minimum node size and maximum number of levels of a tree be pre-specified. Specifying the number of levels of a tree helps restrain the complexity of the subgroups. The number of levels is equivalent to the number of covariates that can define a subgroup. For example, if the maximum number of levels is set to 3, then a subgroup can, at most, be defined by 3 covariates. However, in addition, SIDES also requires the pre-specification of the best number of splits, say M , to be considered for each parent node at each level. For example, if $M=3$, then for any node, all splits will be evaluated for each covariate retaining only the single best split for each covariate. The best splits are then ordered from smallest to largest and the best three covariate splits are chosen to split on. The SIDES procedure also requires the user to either specify the relative improvement parameters to be used, or specify the grid space to be searched for the optimal relative improvement parameter combination

using cross-validation as described earlier. Once the aforementioned parameters have been pre-specified, the SIDES procedure algorithm can be applied as follows

- Initialise: Start at the root node (level 0) consisting of the entire dataset
- Iteration:
 - Step 1 - Evaluate the splitting criterion for all splits of every covariate (exclude covariates already used to define the parent node) retaining only the best split for each covariate. Order the covariates from smallest adjusted p-value to largest adjusted p-value where the adjusted p-values are computed using the Sidak-based multiplicity adjustment. Ordering the adjusted p-values ensures that covariates with a larger number of splits are not favoured i.e. not placed higher up in the ordering due to multiplicity.
 - Step 2 - Select the best M covariates from the ordered best splits. For each of the M splits, form the split creating two child nodes and retain the child node with the larger positive treatment effect, provided it satisfies the relative improvement parameter condition (see equation 6.11). The retained nodes now become parent nodes for the next iteration.
 - Step 3 – Repeat steps 1 and 2 for the newly formed parent nodes
 - Step 4 – Continue the above three steps until either the maximum number of levels is reached or if no more splits can be formed i.e. relative improvement parameter condition not satisfied. In both cases, the previously formed parent nodes become terminal nodes.

The above algorithm is implemented until a SIDES tree has been grown. The resultant tree is a collection of candidate subgroups that all have an enhanced treatment effect

where each subgroup is defined by up to M covariates. However, many of the candidate subgroups might be spurious findings due to inflated type I error rates. Thus to control for this and to remove possibly spurious subgroups, the authors propose a resampling based procedure to adjust the p-value for each candidate subgroup and control the type I error in the weak sense (135). The resampling procedure firstly randomly permutes the rows of the outcome variable and treatment variable in the dataset to form a null dataset. The outcome and treatment variables are kept together and permuted then reattached to the covariate values. It is done in this way to ensure the overall treatment effect is maintained as well as the correlation structure of the covariates. The SIDES procedure is then applied to the null dataset using the same parameters used to grow the initial tree i.e. the same complexity parameters, minimum node size, etc, and the p-value of the best subgroup recorded. This process is repeated many times, e.g. 500 or 1000 times, forming a distribution of p-values. The observed p-values from each of the candidate subgroups is then adjusted by calculating the proportion of p-values in the distribution from the resampling procedure that fall below the observed p-value. More precisely, the adjusted p-value for a given candidate subgroup is computed by

$$adjusted\ p - value = \frac{\text{number of } p - \text{values from null less than observed } p - \text{value}}{\text{number of permutations}}.$$

Once the p-values for each of the candidate subgroups have been adjusted using the resampling procedure, inferences can be made by observing the adjusted p-values and thus the final subgroups chosen. This basically means that the adjusted p-value is a comparator for the unadjusted p-value in the sense that it reflects how true the subgroup found may be. If the unadjusted p-value is significant and the adjusted p-value is non-significant, then this suggests that the identified subgroup is quite possibly a spurious finding.

SIDES algorithm illustration

A simple example will now be considered to better illustrate how the SIDES procedure works; see Figure 6.3. Assuming we have three covariates, say X_1 , X_2 and X_3 , and that we allow SIDES to split up to a maximum of two levels i.e. any candidate subgroups found can be defined by up to two covariates. Moreover, assume that we select the best two splits ($M=2$) at each level of the tree. The SIDES procedure initially starts at level 0 with a single parent node consisting of the entire dataset, illustrated by a black circle. From the first iteration, the covariates X_1 and X_2 contain the best two splits at split points S_1 and S_2 respectively. The treatment effect in subgroup $X_1 \leq S_1$ is substantially larger than the effect in $X_1 > S_1$, thus the sample $X_1 > S_1$ is disregarded and the sample with $X_1 \leq S_1$ is retained thus forming a parent node at level 1. Similarly for the second split on X_2 , the sample $X_2 > S_2$ is retained and the sample $X_2 \leq S_2$ is disregarded thus the sample $X_2 > S_2$ forms another parent node at level 1. Any disregarded nodes after a split has been made have been coloured grey in figure 6.3. Therefore at level 1, there are two new parent nodes (two black circles) that can both be searched for the best two splits respectively in the next iteration. Note that only those covariates that have not already been used to define the parent node are searched for the next split. For example, if at level 1 the covariate $X_1 \leq S_1$ is used to define the parent node, then the next iteration of SIDES can only search covariates X_2 and X_3 for the next best two splits. Exactly the same process is carried out on the two parent nodes at level 1. The first parent node at level 1 can be split by the covariates X_2 and X_3 at split points S_{12} and S_{13} respectively. From these splits, $X_2 \leq S_{12}$ and $X_3 \leq S_{13}$ are retained to form two new parent nodes at level 2. Similarly, the second parent node at level 1 can be split by the covariates X_1 and X_3 at split points S_{21} and S_{23} respectively. From these splits, $X_1 \leq S_{21}$ and $X_3 \leq S_{23}$ are retained to form two more parent nodes at level 2. Thus there are four new parent nodes formed at level 2. If we recall earlier, we specified the maximum number of covariates used to define a subgroup as being two i.e. maximum of two

levels. Thus, the SIDES procedure stops here and the four parent nodes identified at level 2 thus become terminal nodes. Therefore at the end of the procedure, four candidate subgroups have been identified; subgroup1 $\{X_1 \leq S_1 \text{ and } X_2 \leq S_{12}\}$; subgroup2 $\{X_1 \leq S_1 \text{ and } X_3 \leq S_{13}\}$; subgroup3 $\{X_2 > S_2 \text{ and } X_1 \leq S_{21}\}$; subgroup4 $\{X_2 > S_2 \text{ and } X_3 \leq S_{23}\}$. Though this example has been illustrated using Figure 6.3 in the form of a single tree (i.e. in the same form illustrated by the authors) it can also be illustrated as two separate trees defined by the two branches stemming from the root node. The reason we can do this is because each branch stemming from the root node is essentially a different way in which we can initiate the tree growing process starting with the entire dataset. Thus, if we specify that we want the procedure to consider the best M splits for each node, then this means that we can illustrate the final identified subgroup by up to M separate trees with the root node being the entire dataset.

6.4 Discussion

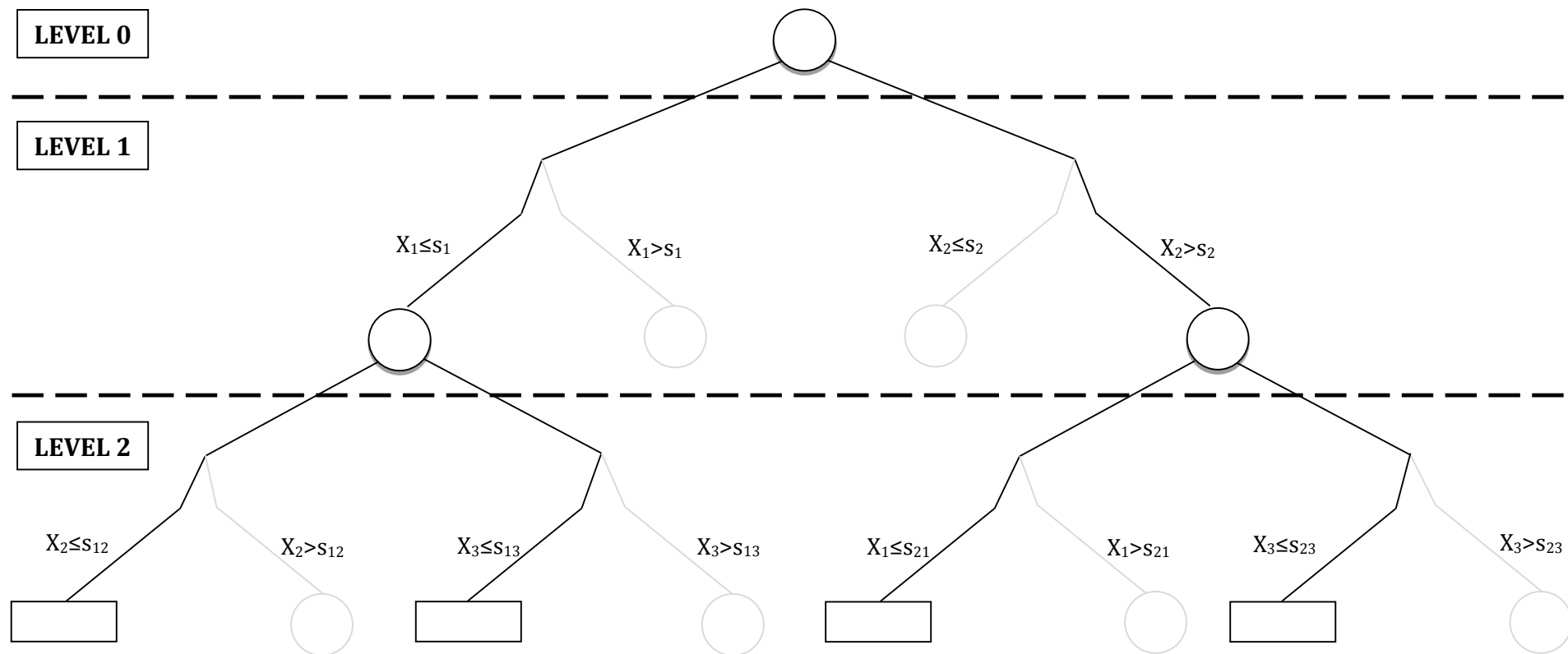
Tree based methods are a promising alternative to performing exploratory subgroup analyses in randomised controlled trials. These methods overcome many of the issues with existing conventional subgroup analyses as highlighted in previous chapters. Moreover, all tree based methods utilise a simple technique referred to as recursive partitioning. Therefore, this chapter provided a detailed description of the recursive partitioning methodology to better understand the method. The method was described in reference to the commonly used and well established CART type procedure simply to demonstrate how the method works. The CART method is typically applied to data where the goal is to detect interactions that are predictive of outcome; however it does not detect treatment effect heterogeneity i.e. treatment-covariate interactions; which is the focus of this thesis. However, there are a number of advanced variants based on the CART type procedure, namely the IT, STIMA and SIDES methods, which do look for treatment effect heterogeneity. These methods were thus described with reference to

the recursive partitioning methodology described earlier on in the chapter since the steps in the algorithm of these methods are relatively similar.

This chapter highlighted a clear difference in the aims of the IT and STIMA methods compared to the SIDES method. The IT and STIMA methods both look to identify subgroups with the aim of maximising the treatment-covariate interaction effect whereas the SIDES method aims to detect subgroups of individuals who have an enhanced treatment effect. Despite there being this difference, from a clinical perspective, both aims are very important.

This chapter gave a detailed insight into tree based methods and the underlying methodology. Moreover, three recently proposed advanced variants for performing subgroup analyses or subgroup identification were also described. Having identified and described these methods, it is important to evaluate them in a variety of simulated scenarios to assess if they actually do what they aim to do. Therefore, the next chapter will describe a simulation study to evaluate these methods with the aim of identifying the best method(s) for performing subgroup analyses or subgroup identification.

Figure 6.3 – Example of the SIDES procedure with two levels



Chapter 7

Simulation study to evaluate tree based methods

7.1 Introduction

The previous chapter introduced the recursive partitioning methodology; a key component of tree based approaches. Although there are several variants of the tree based method that employ a CART type routine, there are only a few that specifically look to identify differential subgroup effects, which is what is of interest in this PhD. Moreover, these methods enable us to identify subgroups defined by multiple baseline characteristics. The tree based methods that do identify differential subgroup effects are interaction trees (IT), Simultaneous Threshold Interaction Modelling Algorithm (STIMA) and Subgroup Identification based on a Differential Effect Search (SIDES) (133, 135, 136). As highlighted in the previous chapter, there is a clear distinction in the aims of both the IT and STIMA methods compared to the SIDES method. The IT and STIMA methods aim to identify baseline characteristics that are moderators of treatment effect i.e. each split in the tree represents a detected interaction effect,

whereas the SIDES method aims to identify subgroups with enhanced treatment effect i.e. each split in the tree represents an identified subpopulation with a large treatment effect.

Both of the aforementioned aims from a clinical perspective are very important. Therefore, all three of these methods stand as good candidate methods that could potentially be extended to an IPD subgroup meta-analysis setting to identify subgroups defined by multiple characteristics. Before considering ways to extend these methods, it is initially important to evaluate how well these methods perform in a single trial setting. This chapter will therefore describe the setup of the simulation study in a single trial setting along with a description of the simple scenarios to be considered. The results of the simulation study will then be presented and assessed to see how well the IT, STIMA and SIDES methods perform in a number of simple simulated scenarios.

7.2 Simulation study setup

Simulation study design

The purpose of this simulation study was to evaluate the performance of the IT and STIMA methods in detecting treatment-covariate interactions and to evaluate the performance of the SIDES method in detecting subpopulations with enhanced treatment effect. In this first instance, we will keep the simulation study relatively simple rather than looking at absolutely everything to ease the assessment of these methods in a single trial setting. Data were simulated using a single linear regression model of the following form,

$$Y = \beta_0 + \beta_1 T + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_1 \cdot X_2 + \beta_5 T \cdot X_1 + \beta_6 T \cdot X_2 + \beta_7 T \cdot X_1 \cdot X_2 + \varepsilon \quad (7.1)$$

where Y is a continuous response variable, T is a treatment variable consisting of two arms (T_1 and T_2), X_1 and X_2 are covariates each with two categories and ε is the

normally distributed error with mean 0 and variance 1. The main effect sizes and interaction effect sizes were thus varied in model (7.1) to create scenarios in which to evaluate the methods. The main effects in the simulation study were set to a constant term, say zero ($\beta_1 = \beta_2 = \beta_3 = 0$), since the methods are not influenced by the main effects i.e. they only detect treatment-covariate interaction effects. Moreover, the coefficients corresponding to the interaction term not including treatment and the two-way interaction were also set to zero i.e. $\beta_4 = 0$ and $\beta_7 = 0$. Using model (7.1), the simulation study evaluated how well these methods perform in three simple scenarios:

- 1) Null model – The null model is just a main effects model with no interaction effects present. This corresponds to model (7.1) with all coefficients set to zero i.e. $\beta_i = 0$. Evaluating the methods in this scenario will allow for the assessment of the type I error rate of these methods.
- 2) Single one-way interaction – This model consists of main effects with the inclusion of a single one-way treatment-covariate interaction. This corresponds to β_5 being specified in the model (7.1).
- 3) Two one-way interactions – This model consists of a main effects model with the inclusion of two one-way treatment-covariate interactions. This corresponds to β_5 and β_6 being specified in the model (7.1).

Three factors were varied in the simulation study; the sample size, the effect size for the $T \cdot X_1$ interaction and the effect size for the $T \cdot X_2$ interaction. To make the entire simulation study more efficient and to mimic all three scenarios using just one model (model (7.1)), a full-factorial design was used. In this full factorial design a set of simulations is performed for every single combination of the three varying factors. Since the simulation study was based on single trial data, a variety of small to large sample sizes were considered; $N=200, 400, 500, 600, 1000$. Standardized interaction effect sizes of 0, 0.2 (small), 0.5 (medium), 0.8 (large) and 1.5 (very large) were

considered for both the $T \cdot X_1$ and the $T \cdot X_2$ interactions, which correspond to the coefficients β_5 and β_6 respectively in the model. The simulated data were set up such that there was an equal proportion of individuals in each category of the treatment variable T , and the two covariates X_1 and X_2 . In total, each combination in the full-factorial design was simulated 1000 times and the results summarized. Altogether there were 125 combinations therefore in total 125,000 simulations were performed for each of the three methods. The simulations were performed using R software 3.0.1.

Data generation procedure

The model given by equation (7.1) was used to generate the data to perform the simulation. The aim was to set up a data generation function such that the required effect sizes can be specified and the data generated. We can represent the model given by equation (7.1) by a $2 \times 2 \times 2$ table of the cell means as shown in table 7.1. Within table 7.1, the means within cells (X_{11}, X_{21}, T_1) , (X_{11}, X_{21}, T_2) , ..., (X_{12}, X_{22}, T_2) are given by a , b , c , d , e , f , g , h respectively. In reality, the number of patients within each of these cells will not always be equal and this needs to be accounted for. For this reason, the marginal means are based on the proportion of imbalance within each covariate, denoted by p , q_1 , q_2 and r (displayed using grey text) in table 7.1 to account for this. The marginal means have been displayed in table 7.1 using bold text. This enables a generalized data generation framework to be setup. Using the tabulation, we can easily formulate general expressions for the marginal mean (β_0), the overall treatment effect (β_1), the overall X_1 effect (β_2), the overall X_2 effect (β_3), the interaction between X_1 and X_2 (β_4), the interaction between X_1 and treatment (β_5), the interaction between X_2 and treatment (β_6) and finally the interaction between treatment, X_1 and X_2 (β_7). These expressions can be written as follows

- $\beta_0 = pq_1ra + pq_1(1-r)b + p(1-q_1)rc + p(1-q_1)(1-r)d + (1-p)q_2re + (1-p)q_2(1-r)f + (1-p)(1-q_2)rg + (1-p)(1-q_2)(1-r)h$
- $\beta_1 = pq_1a - pq_1b + p(1-q_1)c - p(1-q_1)d + (1-p)q_2e - (1-p)q_2f + (1-p)(1-q_2)g - (1-p)(1-q_2)h$
- $\beta_2 = rq_1a + (1-r)q_1b + r(1-q_1)c + (1-r)(1-q_1)d - rq_2e - (1-r)q_2f - r(1-q_2)g - (1-r)(1-q_2)h$
- $\beta_3 = pra + p(1-r)b - prc - p(1-r)d + (1-p)re + (1-p)(1-r)f - (1-p)rg - (1-p)(1-r)h$
- $\beta_4 = ra + (1-r)b - rc - (1-r)d - re - (1-r)f + rg + (1-r)h$
- $\beta_5 = q_1a - q_1b + (1-q_1)c - (1-q_1)d - q_2e + q_2f - (1-q_2)g + (1-q_2)h$
- $\beta_6 = pa - pb - pc + pd + (1-p)e - (1-p)f - (1-p)g + (1-p)h$
- $\beta_7 = a - b - c + d - e + f + g - h$

To find the solutions to the above equations, we simply convert them to matrix form and then invert the matrix. See Appendix C for the matrix form of the above equations. Thus inverting the matrix will provide solutions for computing each of the cell means that are in the same form as that of equation (7.1). The solutions are as follows:

$$a = \beta_0 + (1-r)\beta_1 + (1-p)\beta_2 + (1-q_1)\beta_3 + (pq_1 - q_1 - p + 1)\beta_4 + (pr - r - p + 1)\beta_5 + (q_1r - r - q_1 + 1)\beta_6 + (pq_1 - q_1 - r - p + pr + q_1r - pq_1r + 1)\beta_7$$

$$b = \beta_0 + (-r)\beta_1 + (1-p)\beta_2 + (1-q_1)\beta_3 + (pq_1 - q_1 - p + 1)\beta_4 + (pr - r)\beta_5 + (q_1r - r)\beta_6 + (pr - r + q_1r - pq_1r)\beta_7$$

$$c = \beta_0 + (1-r)\beta_1 + (1-p)\beta_2 + (-q_1)\beta_3 + (pq_1 - q_1)\beta_4 + (pr - r - p + 1)\beta_5 + (q_1r - q_1)\beta_6 + (pq_1 - q_1 + q_1r - pq_1r)\beta_7$$

$$d = \beta_0 + (-r)\beta_1 + (1-p)\beta_2 + (-q_1)\beta_3 + (pq_1 - q_1)\beta_4 + (pr - r)\beta_5 + (q_1r)\beta_6 + (q_1r - pq_1r)\beta_7$$

$$e = \beta_0 + (1-r)\beta_1 + (-p)\beta_2 + (1-q_2)\beta_3 + (pq_2 - p)\beta_4 + (pr - p)\beta_5 + (q_2r - r - q_2 + 1)\beta_6 + (pq_2 - p + pr - pq_2r)\beta_7$$

$$f = \beta_0 + (-r)\beta_1 + (-p)\beta_2 + (1-q_2)\beta_3 + (pq_2 - p)\beta_4 + (pr)\beta_5 + (q_2r - r)\beta_6 + (pr - pq_2r)\beta_7$$

$$g = \beta_0 + (1-r)\beta_1 + (-p)\beta_2 + (-q_2)\beta_3 + (pq_2)\beta_4 + (pr - p)\beta_5 + (q_2r - q_2)\beta_6 + (pq_2 - pq_2r)\beta_7$$

$$h = \beta_0 + (-r)\beta_1 + (-p)\beta_2 + (-q_2)\beta_3 + (pq_2)\beta_4 + (pr)\beta_5 + (q_2r)\beta_6 + (-pq_2r)\beta_7$$

Table 7.1 – 2 x 2 x2 table of within cell means for T, X_1 and X_2

		R		$(1-r)$
		T_1	T_2	
p	X_{11}	a	b	$rq_1a + (1-r)q_1b$
	X_{22} $(1-q_1)$	c	d	$r(1-q_1)c + (1-r)(1-q_1)d$
$(1-p)$	X_{21} q_2	e	f	$rq_2e + (1-r)q_2f$
	X_{22} $(1-q_2)$	g	h	$r(1-q_2)g + (1-r)(1-q_2)h$
		$pq_1a + p(1-q_1)c + (1-p)q_2e + (1-p)(1-q_2)g$		$pq_1b + p(1-q_1)d + (1-p)q_2f + (1-p)(1-q_2)h$

The values for $\beta_0, \beta_1, \beta_2, \beta_3$ and the interaction effects ($\beta_4, \beta_5, \beta_6, \beta_7$), can thus be specified as well as any imbalances we want to consider (p, q_1, q_2 and r) in order to compute the values of each of the cell means (a, b, c, d, e, f, g and h). Outcome data (Y) are then generated for the n individuals within each of the cells such that the mean is equal to the pre-computed cell mean. This is done by simulating the outcome data (Y) from a normal distribution using each of the cell means as the mean, a standard deviation of 1 and specifying n . However, the value of n may vary for each cell if there is imbalance in the proportions. Here we can assume the proportions across treatment arms are the same due to randomization i.e. $r=0.5$. Still, the proportions within the treatment arms can vary i.e. depends on the value of p and q . The simulation setup however considers the simplest setting assuming equal proportions of individuals within each of the cells; thus p, q and r will all be set to 0.5. To consider an example, if we want to simulate outcomes for 40 individuals, then the number of individuals within each of the cells would simply be 5 (40 divided by 8).

Now that the outcome generation procedure has been determined, all that is required is to create a treatment variable, an X_1 variable and an X_2 variable, each with two categories. The values of the categories used when creating these variables e.g. $X_1=0$ or 1, must have specific values to ensure the coefficients estimated by the model are correct. The correct values can be obtained by looking at the term that is multiplied by β_1, β_2 and β_3 respectively in the solution for each of the cells produced having inverted the matrix. For example, those individuals in cell a will have a value of $(1-r)$ for treatment, $(1-p)$ for X_1 and $(1-q_1)$ for X_2 . Thus, the values for the variables depend on p, q and r . Again in this simulation study setup, the proportions are assumed to be equal thus the values for the treatment variable, X_1 and X_2 will be -0.5 or 0.5. In this way, we can generate data to suit our requirements for the simulation study. The data can be easily checked by fitting the regression model in (7.1) to the newly generated data.

Parameter specifications for tree methods

All of the tree methods require some parameters to be specified prior to applying the algorithm to the simulated data. This section will therefore describe the parameter specifications required for the tree methods for estimation. The minimum number of individuals in any node at any given time was set to 10% of total sample size. The maximum number of levels or splits of the fully grown tree produced by the IT and STIMA methods was set to three. The maximum number of levels for SIDES was set to two since there are only two covariates to consider; thus any identified subgroups are only ever defined by up to two covariates. In addition, the number of best splits to consider for each node was also set to two for the SIDES procedure. The STIMA method was set up such that the first split of the tree was forced on the treatment variable in order to detect treatment-covariate interactions. Where the methods require the use of V-fold cross-validation, the number of folds to be used was set to five i.e. 5-fold cross-validation. However for STIMA, as the authors suggest, the 5-fold validation was repeated five times to obtain more stable estimates.

7.3 Final trees grown by methods

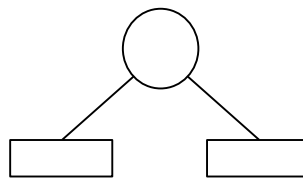
Prior to conducting the simulation study, it is useful to establish what the final correct tree should be for IT, STIMA and SIDES in each of the three scenarios. More specifically, the correct tree for each method in each scenario is determined by the size of the final tree selected and the covariates used to form the final tree. Thus the correct final trees for each method will now be described.

Final trees grown by IT method

Scenario 1 – In scenario 1 (null model), the final tree selected by the IT procedure should consist of just a single node i.e. the root node, that consists of the entire dataset.

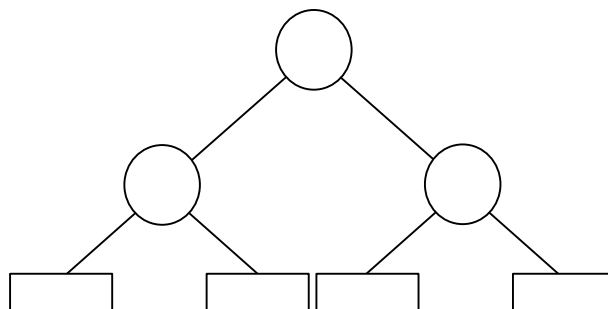
Scenario 2 – In scenario 2 (single one-way interaction), the final tree selected by the IT procedure should be of size 2 i.e. it has two terminal nodes, see figure 7.1. Moreover, the covariate used to form the single split should be either X_1 or X_2 depending on which of the two has been set up in the simulation as having an interaction effect with treatment.

Figure 7.1 – Final tree produced by the IT method for a single one-way interaction



Scenario 3 – In scenario 3 (two single one-way interactions), the IT procedure should select a final tree of size 4 i.e. it has four terminal nodes, see figure 7.2. The final tree should select both X_1 and X_2 as the covariates used to form the two splits in the tree.

Figure 7.2 – Final tree produced by the IT method for two single one-way interactions

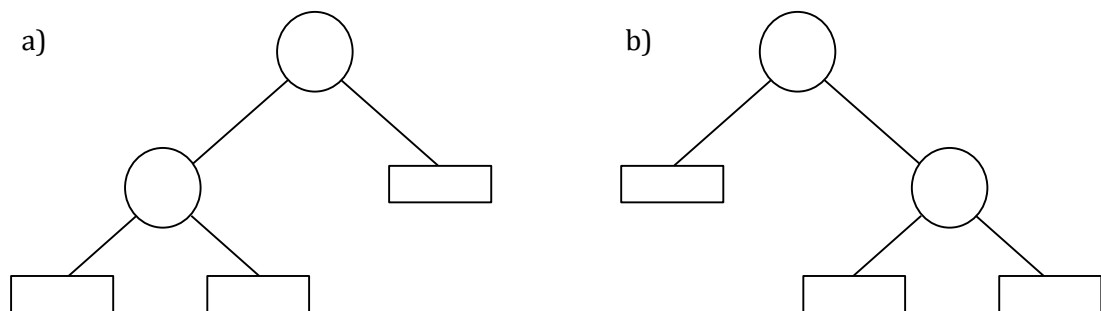


Final trees grown by STIMA method

Scenario 1 – In scenario 1 (null model), the final tree selected by the STIMA procedure should consist of just a single node i.e. the root node, that consists of the entire dataset. Moreover, the linear regression model estimated by STIMA should consist of main effects only.

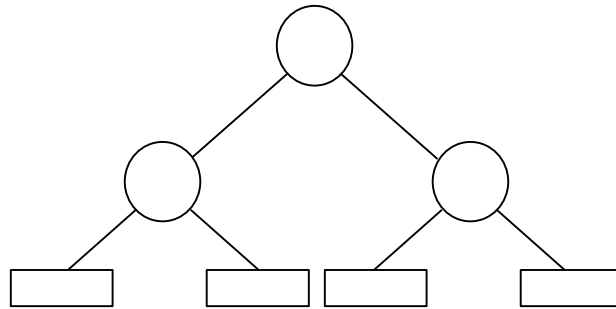
Scenario 2 – In scenario 2 (single one-way interaction) the final tree produced by the STIMA procedure should be of size 3 i.e. consists of three terminal nodes. Recall that the first split is forced on treatment; therefore a single one-way interaction will require a single split on either one of the two nodes formed from the forced split as displayed in figure 7.3. Recall from the previous chapter, when a split is found by STIMA, it is added to the model in the form of an indicator variable and the model is re-estimated and updated thus incorporating the detected interaction effect. Therefore, of the nodes formed from the first forced split on treatment, it doesn't matter which one of the two nodes it splits on afterwards as the model will be updated accordingly.

Figure 7.3 – Final tree produced by the STIMA method for a single one-way interaction



Scenario 3 – In scenario 3, the final tree produced by the STIMA procedure should be of size 4 i.e. consists of four terminal nodes. The root node is first split by treatment to form two child nodes. Each of the child nodes is then split by one of the covariates X_1 or X_2 to form the final tree, see figure 7.4. Again with each iteration, the STIMA method updates the model to include the detected interaction.

Figure 7.4 – Final tree produced by the STIMA method for two single one-way interactions

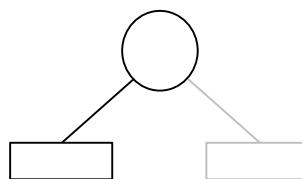


Final trees grown by SIDES method

Scenario 1 – In scenario 1 (null model), the final tree selected by the SIDES procedure should consist of just a single node i.e. the root node, that consists of the entire dataset.

Scenario 2 – In scenario 2 (single one-way interaction), the SIDES procedure identifies a single candidate subgroup defined by a single covariate, either X_1 or X_2 , depending on which covariate the simulation has set up to have an interaction with treatment. Recall that having identified a best split, the SIDES procedure retains the node or subgroup with the larger positive treatment effect. Essentially, this is exactly the same as the tree produced by the IT procedure in figure 7.1, but then the node with the smaller treatment effect is removed; thus identifying a subgroup or subpopulation with an enhanced treatment effect. The final tree produced by SIDES in this scenario is displayed in figure 7.5 where the removed node has been shaded grey.

Figure 7.5 – Final tree produced by the SIDES method for a single one-way interaction



Scenario 3 – In scenario 3 (two single one-way interactions), the SIDES procedure produces two candidate subgroups of enhanced treatment effect; however the structure of the tree for the two candidate subgroups can vary. The first way SIDES can identify the two regions with large positive treatment effect is by identifying two candidate subgroups at level 1 where one subgroup is defined by splitting on X_1 and the other subgroup is defined by splitting on X_2 , as shown in final tree (A) in figure 7.6. As mentioned in the previous chapter, the tree in figure 7.6 can also be illustrated as two separate trees as shown in figure 7.7.

Figure 7.6 – Final tree (A) produced by the SIDES method for two single one-way interactions; illustrated using a single tree.

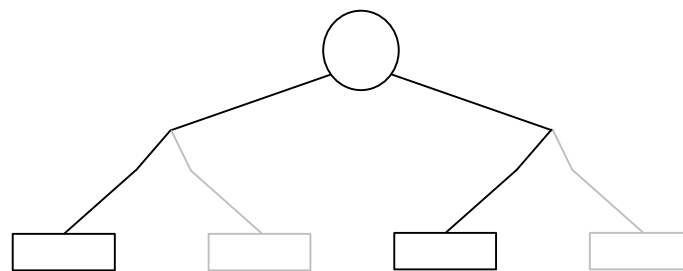


Figure 7.7 – Final tree (A) produced by the SIDES method for two single one-way interactions; illustrated using two separate trees.



Another way SIDES can identify the two regions is by further splitting the candidate subgroups retained at level 1 using the covariate that has not already been used to define that particular subgroup, as shown in final tree (B) in figure 7.8. For example, if a candidate subgroup has been split using X_1 at level 1, then at level 2 the procedure will go on to form a split using X_2 , and vice versa. Figure 7.8 shows only one of the nodes at level 1 being split further to form a subgroup at level 2. However, it is also

possible for the other subgroup at level 1 to be split further to form a subgroup at level two as well. Again, the tree illustrated in figure 7.8 can also be illustrated using two separate trees as shown in figure 7.9.

Figure 7.8 – Final tree (B) produced by the SIDES method for two single one-way interactions; illustrated using a single tree.

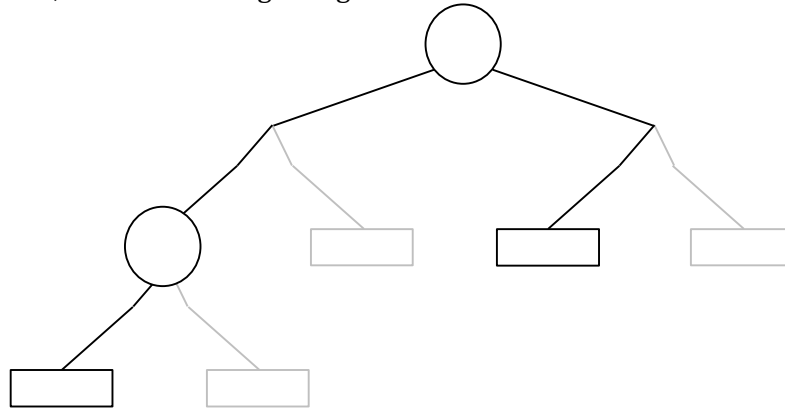
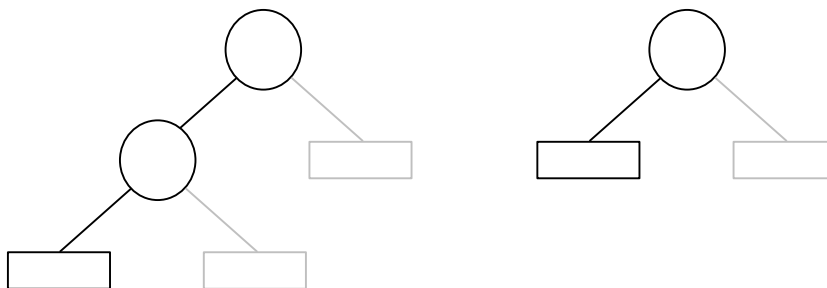


Figure 7.9 – Final tree (B) produced by the SIDES method for two single one-way interactions; illustrated using two separate trees.



7.4 Simulation Results

IT method results

The results from the IT procedure simulation are displayed in table 7.2. The results from scenario 1 (the null model) are reflective of the type I error rate for the IT procedure. More than 90% of the time, the IT procedure correctly selects a tree of size 1, i.e. the root node only, when there are no interaction effects present. When only a

single one-way interaction is present (scenario 2), the IT procedure does not perform well when the standardized interaction effect size is small. It performs reasonably well, with over 70% correct trees grown, when detecting a medium interaction effect size provided the total sample size is ≥ 600 . The method performs quite well and detects the correct tree the majority of the time when the sample size is ≥ 400 when detecting a large interaction effect size and for all sample sizes ≥ 200 when detecting a very large interaction effect size. In scenario 3 where there are two one-way interactions, it is apparent that the performance of the IT method is affected in the presence of more than a single one-way interaction effect. If one of the two one-way interactions has either a small or medium effect size, then the method does not perform well. A closer inspection of the results from individual simulations suggests that when one of the interaction effects is either small or medium and the other interaction effect is greater i.e. large or very large, then the IT method seems to only detect the larger effect. When one of the interaction effects is large and the second interaction effect is either large or very large then the method detects the correct tree the majority of the time provided that total sample size is ≥ 600 . The method works very well when both interaction effects are very large for all sample sizes ≥ 200 .

STIMA method results

The simulation results from applying the STIMA procedure are displayed in table 7.3. In scenario 1 where no interaction effects are present, the STIMA approach always identifies the correct tree consisting of a single root node. In other words, outcome is best predicted using a linear combination of the main effects only. When a single one-way interaction is present, the method detects the correct tree the majority of the time if the effect size is large and the sample size is 1000, or for all sample sizes ≥ 200 when the effect size is very large. In scenario 3 where two one-way interactions are present, the STIMA method does not perform well for small and medium effect sizes. The

method detects the interactions a majority of the time when the sample size is 1000 and either both interactions are large or when one is large and the other is very large. The method also performs well when both the interactions are very large and the sample size is ≥ 200 . A closer inspection of the results indicates that interactions are hardly ever detected when both the interaction effects are small. Moreover, when the interaction effect sizes differ, the STIMA method detects the larger interaction effect the majority of the time. When observing all of the results for STIMA, there seems to be something rather peculiar happening; the performance consistently deteriorates when considering a sample size of 500. This will be reviewed later on in the discussion section.

SIDES method results

The simulation results from applying the SIDES procedure are displayed in table 7.4. In scenario 1 where no subgroups with enhanced treatment effect are present i.e. no interaction effects, the SIDES procedure correctly identifies no candidate subgroups approximately 90% of the time. In scenario 2 where there is a single subgroup with enhanced treatment effect (a single one-way interaction), the SIDES method does not perform so well when the effect size is small. When there is a medium sized interaction effect, the method detects the correct subgroup around 80% of the time provided the total sample size is about 1000. When the effect size is large, the method detects the correct subgroup a majority of the time if the total sample size is > 400 . For very large effect sizes it detects the correct subgroup a majority of the time for all sample sizes ≥ 200 . In scenario 3 where there are two regions with enhanced treatment effect i.e. two one-way interactions present, the SIDES procedure does not perform very well if one of the two interaction effects is small. A closer inspection of the individual simulation results suggests that when one effect size is small and the other effect size is greater, then the SIDES procedure only detects the larger effect size; thus identifying

just one candidate subgroup each time. When both the interactions have a medium effect size, or if one has a medium effect and one has a large effect then the method detects the correct subgroups a majority of the time for a sample size of 1000. If the two interactions are large and very large respectively, then the method performs well provided the sample size is ≥ 400 . When both interaction effects are very large, the method seems to work very well in detecting both candidate subgroups with large treatment effects.

7.5 Discussion

A simple simulation study was performed in this chapter to evaluate the performance of the IT and STIMA procedures in detecting interaction effects and to evaluate the SIDES procedure in detecting subpopulations with enhanced treatment effect. The IT and STIMA procedures both have the same objective of detecting subgroups that maximize the interaction effect. Comparison of the results suggests that overall, the IT method performs much better than the STIMA method. In particular, it performs well when detecting large or very large interaction effects. In addition, the SIDES procedure performs quite well in detecting subgroups with large treatment effects.

As highlighted in the results section, there was some peculiarity observed in the STIMA results. In the presence of two one-way interactions, one would expect the performance of STIMA to improve as the sample size increased from 200 to 1000; however, this was not the case. When the sample size is 500, the performance seems to worsen. A component of the STIMA method that might influence this result is the 0.50-SE rule and 0.80-SE rule suggested by the authors for selecting the final tree. The simulations were repeated using the typical CART 1-SE rule however this made no difference to the results. To investigate the performance further, the STIMA simulation

Table 7.2 – Simulation results for the IT method. Results display % of correctly identified final trees.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	200	92.15	6.75	30.80	64.40	89.80
	400	93.10	10.90	55.00	86.20	90.60
	500	93.85	13.35	63.70	87.15	90.90
	600	93.10	16.00	72.05	90.70	90.60
	1000	93.70	22.50	85.65	90.75	91.20
Small=0.2	200	6.40	0.10	0.45	1.10	1.30
	400	10.70	0.40	1.30	1.50	1.85
	500	13.25	0.25	1.70	2.70	2.55
	600	14.45	0.50	2.55	3.35	2.65
	1000	23.65	1.25	5.55	5.30	5.20
Medium=0.5	200	28.80	0.25	1.80	4.65	7.45
	400	54.75	0.95	8.95	21.50	21.85
	500	63.60	1.95	14.85	27.85	31.10
	600	70.40	2.30	21.20	39.20	39.95
	1000	86.00	4.85	52.00	68.45	67.90
Large=0.8	200	66.60	0.60	4.80	15.35	33.65
	400	85.05	1.75	19.85	54.45	68.80
	500	88.10	2.55	30.40	70.90	79.10
	600	88.95	3.15	38.65	81.50	88.70
	1000	90.90	5.55	66.85	97.80	98.95
V. Large=1.5	200	89.60	1.30	8.10	30.60	89.70
	400	91.05	2.05	23.15	69.55	99.95
	500	91.55	1.85	31.75	81.80	100.00
	600	90.75	2.90	37.05	90.00	100.00
	1000	92.45	6.05	67.60	98.80	100.00

Table 7.3 – Simulation results for the STIMA method. Results display % of correctly identified final trees.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	200	100.00	0.10	0.50	8.10	99.90
	400	100.00	0.10	7.00	49.60	100.00
	500	100.00	0.20	6.90	57.70	100.00
	600	100.00	0.00	8.60	69.20	100.00
	1000	100.00	0.00	16.10	90.10	100.00
Small=0.2	200	0.00	0.00	0.00	0.10	0.10
	400	0.20	0.00	0.30	0.10	0.00
	500	0.00	0.00	0.00	0.00	0.00
	600	0.10	0.00	0.10	0.30	0.00
	1000	0.00	0.00	0.40	0.30	0.00
Medium=0.5	200	0.60	0.10	0.40	1.20	1.60
	400	7.50	0.40	6.70	11.90	4.90
	500	6.30	0.10	2.90	7.20	0.20
	600	8.80	0.50	8.20	15.70	7.70
	1000	16.40	0.20	14.50	22.20	15.20
Large=0.8	200	7.20	0.20	2.00	6.70	12.30
	400	48.30	0.60	12.20	45.20	43.90
	500	56.00	0.10	7.90	35.90	18.10
	600	69.20	0.30	16.50	61.80	65.20
	1000	91.50	0.20	20.50	86.40	89.90
V. Large=1.5	200	84.30	0.20	1.60	11.80	78.10
	400	100.00	0.20	5.80	42.70	99.80
	500	100.00	0.00	0.90	19.10	98.70
	600	100.00	0.00	7.10	64.50	100.00
	1000	100.00	0.00	14.50	89.00	100.00

Table 7.4 – Simulation results for the SIDES method. Results display % of correctly identified candidate subgroups.

T^*X_1 Standardized interaction effect size	N	T^*X_2 Standardized interaction effect size				
		None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	200	88.50	8.10	26.50	54.20	92.70
	400	87.80	11.70	47.60	81.00	97.30
	500	89.80	13.80	51.70	85.30	99.20
	600	89.00	13.80	64.40	92.10	98.60
	1000	89.40	20.10	80.50	95.40	98.30
Small=0.2	200	8.80	3.60	8.40	12.00	7.60
	400	12.90	5.80	15.20	18.20	7.90
	500	12.60	8.00	20.30	17.30	8.80
	600	15.70	10.50	21.00	18.50	9.70
	1000	23.40	16.90	30.60	23.70	17.90
Medium=0.5	200	27.20	7.00	18.80	34.20	30.10
	400	48.20	12.00	45.60	57.80	45.50
	500	55.80	15.30	52.50	65.00	52.70
	600	62.50	20.50	66.70	75.40	64.90
	1000	82.50	28.20	87.60	89.10	83.80
Large=0.8	200	54.80	10.50	30.90	57.00	63.80
	400	84.90	15.20	56.00	87.10	89.00
	500	89.90	15.00	63.60	93.90	93.00
	600	95.40	17.30	72.80	97.60	96.50
	1000	97.30	20.30	91.30	99.90	99.80
V. Large=1.5	200	95.40	7.20	23.70	60.30	97.90
	400	98.00	6.90	45.80	86.30	100.00
	500	98.90	7.90	54.70	93.50	100.00
	600	98.80	8.40	62.50	95.90	100.00
	1000	98.70	15.70	84.20	99.50	100.00

was re-run using a variety of sample sizes within the same range (N=200, 400, 416, 456, 488, 500, 528, 560, 584, 600, 616, 640, 664, 696, 1000). The results of this simulation fluctuated across the sample sizes investigated. For example, table 7.5 displays the proportions for the different tree sizes grown by the method when one interaction effect was very large and the other interaction effect was large for the aforementioned sample sizes. In this example, the correct tree identified by the method should be of size four i.e. it has four terminal nodes. Here we can see that the effect of the sample size is non-monotonic in a complicated way for the method.

Table 7.5 – Simulation results presenting the proportion (%) of the different tree sizes obtained for a range of sample sizes to investigate STIMA method when a very large and a large interaction effect are present

N	Tree Size			
	1	2	3	4
200	7.0	0.0	81.2	11.8
400	0.0	0.0	57.3	42.7
416	0.0	0.0	94.3	5.7
456	0.0	0.0	94.6	5.4
488	0.0	0.0	86.5	13.5
500	0.0	0.0	80.9	19.1
528	0.0	0.0	62.9	37.1
560	0.0	0.0	37.6	62.4
584	0.0	0.0	92.2	7.8
600	0.0	0.0	35.5	64.5
616	0.0	0.0	73.0	27.0
640	0.0	0.0	31.5	68.4
664	0.0	0.0	81.1	18.9
696	0.0	0.0	75.5	24.5
1000	0.0	0.0	11.0	89.0

The peculiarities observed in the STIMA results were thoroughly investigated however it was extremely difficult to determine why the method performs in the way it does. Moreover, Dusseldorp et al recently released the STIMA package in R software in November 2013. This code was used to repeat the simulation study and the same

issues were observed. Therefore, this method will not be considered going forward in this thesis since the results clearly suggest that the IT procedure performs better.

The results from the simulation study suggest that the IT procedure is a good approach for detecting large and very large interaction effects in a single trial based setting. In addition, the SIDES procedure is a good approach for detecting subgroups with enhanced treatment effect. Thus both of these procedures are worth developing and extending such that they can be applied to an individual patient data (IPD) subgroup meta-analyses framework. The simulation study in this chapter was based on a single trial based setting. Thus what is even more promising is that the IPD setting will have considerably more data and therefore theoretically this should improve the performance of both IT and SIDES in detecting small and medium interaction effects. The next chapter will therefore detail the proposed extension of the IT and SIDES methods to an IPD subgroup meta-analyses setting. Moreover, a simulation study will be performed to evaluate the performance of the proposed method extensions.

Chapter 8

Extension of recursive partitioning approaches to IPD meta-analysis

8.1 Introduction

The previous chapter described and implemented a simulation study to evaluate the performance of the IT, STIMA and SIDES methods in a single trial setting. The results of the simulation study suggested that the IT approach was far better than the STIMA method in detecting treatment-covariate interactions. Moreover, the SIDES approach also performed well in detecting candidate subgroups with enhanced treatment effect. Therefore, both the IT and SIDES procedures were highlighted as promising approaches that could be extended to an individual patient data (IPD) subgroup meta-analyses setting.

The implementation and evaluation of subgroup analyses thus far in this thesis has been based on a single trial setting. When considering IPD from multiple similar

studies, the statistical analyses become rather more complex due to the hierarchical or clustered data structure. However there are a couple of approaches, namely the two-stage and one-stage approach, that are used for IPD meta-analyses to take into account the data structure when synthesizing or estimating effects (146). These approaches will be worth considering when thinking about the extension of the IT and SIDES methods.

A number of extensions of tree based methods for data that have some correlated structure have been proposed. However, most of these proposed extensions are all in the setting of longitudinal or repeated measures data where responses are collected over time at several time points (147-152). Here we are interested in tree methods for hierarchical or multilevel data structures using fixed or mixed effects. A mixed-effects regression tree has been proposed by Hajjem et al that grows a tree for the fixed component of the model (153). However, the aims of the aforementioned extensions are all to do with determining the best model for predicting response. None of these methods look to specifically identify treatment effect heterogeneity or subgroups with enhanced treatment effect. Hence, this chapter presents extensions to the IT and SIDES methods as novel statistical approaches to specifically identify subgroups in a multilevel or hierarchical data structure setting.

This chapter will initially give a brief introduction to IPD meta-analyses followed by a description of the statistical methods currently used to perform subgroup analyses in this setting. Thereafter, a proposed extension of the IT and SIDES methods in an IPD subgroup meta-analyses framework will then be described. Finally, a simulation study will be performed to evaluate the proposed extension in a number of scenarios and the results presented.

8.2 Introduction to IPD meta-analyses

An IPD meta-analysis uses individual patient data collected from several similar studies. There are several advantages of having IPD compared to having aggregated data. For example, the IPD can be used to replicate the results from the original study, perform analyses adjusting for important covariates and investigate differential treatment effects (71). Though there are several advantages, as with any method, there are also a number of disadvantages. The main disadvantage is that the entire process demands a lot of time, money and effort and is thus resource intensive. In particular, it involves several processes such as creating a data sharing agreement document, contacting authors of relevant studies, setting up a data transfer protocol using a secure link, obtaining the data, cleaning the data, resolving any data queries with data provider and finally ensuring the format of the data is consistent across trials. Despite these resource related disadvantages, the improvement in the quality of the analyses and thus the precision and reliability of the results makes IPD meta-analyses a desired approach for evidence synthesis. In particular, especially in relation to the work in this thesis, it is an ideal framework for exploring modifiers of treatment effect. However, if performing an IPD meta-analysis is beyond ones capacity and resources, then one can always employ a meta-analysis approach using aggregate data as an alternative.

A key fact about any meta-analysis study, whether it is using IPD or aggregate data, is that the quality of the analyses is totally dependent on the quality of the studies used. If the studies used are of a poor design and thus of a poor quality, then the meta-analyses will also be of a poor quality. Therefore, it is recommended that a quality assessment is made on the original studies considered for inclusion in the IPD meta-analyses (71).

8.3 Statistical methods for IPD subgroup meta-analyses

When performing IPD meta-analyses, it is very important to recognize the hierarchical structure or clustering and incorporate it into the statistical modelling. In fact, it is well recognized that any study with an underlying hierarchical or clustered structure should use an appropriate statistical method to account for the clustering (154, 155). For example, each individual in an IPD meta-analysis of LBP studies will be associated with a particular study in the pooled dataset. Thus in a multi-level model, the individuals form the first level in the hierarchy (level 1) and the studies form the second level (level 2). This hierarchical structure implies that individuals selected randomly from a single particular study will be more similar than individuals randomly selected from several studies; thus introducing between-study variation. Simply ignoring the clustering of patients within trials during analyses is inappropriate. In a standard linear model, the clustering is ignored and it is assumed that the individual observations are independent; hence the error values for each observation will also be unrelated. As we know, due to clustering at the study level, the outcomes within each study will have some degree of correlation. If the outcomes are correlated, the error terms will also be correlated thus violating the independence assumption of the linear regression model. Therefore, a linear regression model applied to IPD will not be able to provide reliable estimates of the coefficient standard errors. The estimated standard errors will be underestimated and this could potentially lead to false inferences claiming real effects exist when in actual fact they don't. For that reason it is very important, as demonstrated in a recent study, that the clustering of patients within studies is accounted for when performing IPD meta-analyses (156). For this reason IPD meta-analyses typically use either a two-stage approach or a one-stage approach to account for the clustering (146). A two-stage approach conducts the analysis using conventional linear regression for each study separately and then synthesizes the

results using well established meta-analysis techniques. A one-stage approach on the other hand uses a mixed-effect model, also referred to as a multilevel model or hierarchical model, to fit a single model to the pooled IPD. It is called a mixed effect model because the model consists of fixed effects and random effects; where the random effects are used to capture the variation at different levels. A fixed-effects model including indicator variables for each study can also be applied for the one-stage approach where all the components of the model are fixed.

There are several papers available detailing the application and advantages of using a two-stage and a one-stage approach (157-159). However, when considering the extension of tree based methods to an IPD setting, it would very difficult and computationally intensive to use a two-stage approach. There are two ways in which the two-stage approach can be implemented. One approach would be to naively grow a tree for each trial separately; however each trial will probably grow a different tree thus making it impossible to synthesize the results. Another approach would be to evaluate every split for each covariate using the splitting function for each trial separately and then synthesize the score across the trials using some weighted average (as done in aggregate data meta-analyses). However, a danger with this is that if one of the trials does not contain the value of the split being considered, then a score will not be computed for that trial and thus the information from that trial will be lost. For example, if a tree method was considering a split on gender (males vs. females) in each trial separately and if one trial had just females in it, then no score would be computed and so the trial would not contribute anything to the estimation of the effect i.e. loss of information. Moreover, such a procedure would be computationally intensive. The one-stage approach on the other hand would not experience the aforementioned difficulties associated with the two-stage approach. Hence, a one-stage approach is better suited to tree based methods and their application. The main advantage of the two-stage method

is simplicity, but as it is lost here, there is no reason to pursue it. Therefore, only the one-stage approach will be considered going forward.

In a one-stage approach, the covariates in the mixed effect model can be set-up to have fixed effects or random effects to account for the clustering. To account for the clustering, one approach would be to use a standard linear regression model and add indicator variables to the model for each study (fixed effects) as follows:

$$Y_{ij} = \beta_{0_i} + \beta_1 T_{ij} + \beta_2 X_{ij} + \beta_3 T_{ij} \cdot X_{ij} + \varepsilon_{ij} \quad (8.1)$$

where β_{0_i} is a vector of indicator variables for each study, the ij subscript denotes the i -th observation in the j -th study and ε_{ij} is the normally distributed error term. The model therefore allows each study to have a different intercept and is referred to as the fixed-effects model. This is basically a general linear model that adjusts for the trial effects by including them as indicators in the model. Instead of adding fixed-effects for trials as shown in equation (8.1), another option for a one-stage approach would be to set the study level covariate as having a random-effect. This basically means that the equation is of the same form but the β_{0_i} term in the model is assumed to be normally distributed with mean β_0 and variance $\sigma_{\beta_0}^2$. Thus the fully specified model can be written:

$$Y_{ij} = \beta_{0_i} + \beta_1 T_{ij} + \beta_2 X_{ij} + \beta_3 T_{ij} \cdot X_{ij} + \varepsilon_{ij} \quad (8.2)$$

$$\beta_{0_i} \sim N(\beta_0, \sigma_{\beta_0}^2)$$

The models specified in equations (8.1) and (8.2) are referred to as the fixed-effects model and random-effects model respectively. For both of these models, the intercepts differ for each study however the slopes remain the same.

It is also possible that covariates may differ across studies and so this also needs to be accounted for when using either fixed-effects or random-effects models. For example if the treatment effect is different across studies, then this could be accounted for in a fixed-effect model by including a treatment by study interaction term. In a random-effects model, random effects can be placed on the treatment variable to give the following model

$$Y_{ij} = \beta_{0i} + \beta_{1i}T_{ij} + \beta_2X_{ij} + \beta_3T_{ij} \cdot X_{ij} + \varepsilon_{ij} \quad (8.3)$$

$$\beta_{0i} \sim N(\beta_0, \sigma_{\beta_0}^2)$$

$$\beta_{1i} \sim N(\beta_1, \sigma_{\beta_1}^2)$$

By doing so, the model will have a random intercept and a random treatment effect. In this manner, as illustrated by the example models specified thus far, mixed effects models can be fitted to best incorporate the correlations inherent within the hierarchical data structure to obtain reliable parameter estimates.

Parameter estimation

This section provides a very brief overview as to how the commonly used REML approach is used for parameter estimation in mixed-effects modelling. For a more detailed description, one can refer to Pinheiro et al (160).

The one-stage mixed-effect models (equations (8.2) and (8.3)) make use of maximum likelihood (ML) or restricted (or residual) maximum likelihood (REML) to obtain parameter estimates. Of the two, the REML approach to estimation is preferred as it provides unbiased estimates of the variance parameters and performs well when the data is unbalanced (160, 161). The REML approach works by maximizing the likelihood

for the two components of the mixed model i.e. the fixed effects component and the random effects component. We can write the two components of a mixed model in a general matrix form as follows

$$Y = XB + ZU + e$$

where Y is a $N \times 1$ vector of the reported outcomes, X and Z are the covariate matrices for the fixed component and the random component respectively, B and U are both $N \times 1$ vectors containing the fixed effect coefficients and the random effects respectively and finally e is vector that consists of the residuals. Typically, both U and e have a multivariate normal distribution (MVN) of the form

$$U \sim MVN(0, H)$$

$$e \sim MVN(0, R)$$

Initially the covariance components H and R i.e. \hat{H} and \hat{R} , are estimated using REML so that the parameters B and U can be estimated thereafter. As a whole, the model has a MVN distribution of the form $Y \sim MVN(XB, ZHZ' + R)$ where B is estimated by $(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y$ where $\hat{V} = Z\hat{H}Z' + \hat{R}$, and the U component is estimated using a shrinkage estimate of the form $\hat{H}Z'\hat{V}^{-1}(Y - X\hat{B})$ (157, 160). The estimates of the covariance H for the random component are referred to as REML estimates as it estimates the proportion of the variance explained by the between-study heterogeneity.

8.4 Proposed extension of tree methods

To recap, the previous sections in this chapter have so far introduced the IPD meta-analyses framework and discussed the statistical issues with applying conventional analyses in this setting. Moreover, statistical approaches currently used to better deal

with the clustering of individuals within studies, in particular the one-stage fixed-effects and mixed-effects modelling approaches, were introduced. Specifically, a number of simple fixed-effect and mixed models were described that can be used to perform subgroup analyses; fixed intercept model, random intercept model and the random intercept and random slope model.

The objective of this section of the chapter is to propose extensions to the IT and SIDES methods such that they can be applied to IPD to perform subgroup analyses. A natural approach to consider is to somehow incorporate the currently used fixed-effects and mixed-effects models into the tree based procedures to ensure they account for the clustering within studies. A recent critical review of statistical methods for detecting interactions in IPD meta-analyses by Fisher et al proposed a model that was considered to be the most basic yet useful for a one-stage approach. This model includes a vector of study indicator variables i.e. β_{0i} (fixed effects) with fixed covariate and interaction effects i.e. β_2 and β_3 respectively have fixed effects, and with the treatment effect either being fixed, as shown in equation (8.1), or random, as shown in equation (8.3) (162). The authors also note that any of the components in this basic model can potentially be specified as being either fixed or random. Therefore, a number of variations of this basic model can be considered i.e. varying the fixed and random effects components, when considering possible extensions. At this point, let's assume that the intercept can be either fixed or random and that the treatment, covariate and the interaction are the same for all trials i.e. have fixed effects. Hence there are three basic model variations (including a null model) that can be considered

- Model A – a null model that ignores the clustering i.e. a general linear model that does not include fixed or random trial effects

- Model B – a model with fixed trial effect, fixed treatment effect and fixed interaction effect (Fixed-effects model) – as shown in equation (8.1)
- Model C – a model with random trial effect, fixed treatment effect and fixed interaction effect (Mixed-effects model) – as shown in equation (8.2)

Having discussed the possible basic models that can be considered, we then need to think about how they can be incorporated into extending the tree methods for application to IPD subgroup analyses. It is important to note at this stage that the aim of the extended IT and SIDES methods are not to obtain some pooled effect size estimate e.g. treatment effect or interaction effect, which is typically the aim of most IPD meta-analyses. Rather, the aim will be to identify subgroups of individuals defined by multiple characteristics that either maximize the differential treatment effect (as done by IT) or have enhanced treatment effect (as done by SIDES).

IT method extension (IPD-IT)

Recall that both IT and SIDES use recursive partitioning which heavily relies on a splitting criterion. Therefore, in order for both IT and SIDES to be extended to an IPD setting, they require a new splitting criterion to be defined. Let us first consider the extension of the IT method. In a single trial setting, the IT procedure used a splitting criterion based on the t-test statistic (Chapter 6 – equation (6.8)) to test the null hypothesis that the coefficient of the interaction effect is equal to zero. Therefore, a natural extension to an IPD meta-analysis setting is to use a splitting criterion based on the t-test statistic for the interaction effect when fitting a fixed-effects or mixed-effects model. The model would have accounted for the correlation within studies and so the estimate of the standard error for the interaction effect would be reliable; hence the estimate of the t-test statistic will be reliable. The t-test statistic is computed by

$$t(s) = \frac{\hat{\beta}_3}{\sigma_{\beta_3}}$$

where $\hat{\beta}_3$ is the estimated coefficient of the interaction term and σ_{β_3} is the associated estimated standard error. The computed statistic can either be positive or negative, indicating the direction of the interaction. The tree growing procedure aims to maximize the interaction effect at each level of the tree regardless of the direction of the interaction. For example, it could either be a large positive interaction or a large negative interaction. Hence a splitting criterion is required that focuses on maximizing the absolute interaction effect. Thus, the splitting criterion can be defined as

$$G(s) = t^2(s)$$

which is analogous to the splitting criterion used for the IT procedure in a single trial scenario. The criterion $G(s)$ is essentially the Wald test statistic i.e. $\frac{\hat{\beta}_3^2}{\sigma_{\beta_3}^2}$ to test the null hypothesis that the interaction effect is equal to zero. Therefore, during the tree growing process, an optimal split is defined as the split that maximizes $G(s)$. The splitting criterion can be estimated using either a fixed-effects model (model B) or a mixed-effects model (model C). Note here that the splitting criterion associated with model A is computed in the same way as the original IT method for a single trial. What we now require is to choose which out of these three models (A, B and C) provides the best estimate of $G(s)$ having accounted for the hierarchical data structure so that the IT method grows the correct size tree. This decision depends on what you believe about the distribution of the data. Therefore, it would be good to compare the three models as estimators of $G(s)$ to see how badly the simpler models do when the data really come from a more complex model. This will be evaluated in the next section of this chapter by performing a simulation study. From now onwards, the extended IT method

will be referred to as IPD-IT, which refers to three methods corresponding to models A, B, and C above.

SIDES extension (IPD-SIDES)

A new proposed splitting criterion for the IT method in an IPD meta-analysis setting has now been defined (IPD-IT). Let us now consider the extension of the SIDES method to an IPD framework. Recall that the SIDES procedure in a single trial setting simply computes the one-sided t-test statistic for treatment effect in both the child nodes separately. The procedure also stores the associated p-value from the one-sided test for inferential purposes should the node be chosen to define a subgroup. Thereafter, these two statistics are fed into the splitting criterion (Chapter 6 – equation (6.10)), which is essentially a test for interaction, to compute a p-value to test the absolute value of treatment effect heterogeneity between the two nodes. Due to asymptotics, the t-test statistics are treated as z-statistics and thus the splitting criterion computes a p-value assuming the test statistics have a standard normal distribution. Hence, one can easily extend the SIDES method by computing the p-value in a similar manner using the same splitting criterion but substituting in the treatment effect t-test statistic

$$t(s) = \frac{\hat{\beta}_1}{\sigma_{\beta_1}}$$

from either the fixed-effect model (Model B) or the mixed-effect model (Model C), where $\hat{\beta}_1$ is the treatment effect estimate and σ_{β_1} is the associated standard error. Though we can obtain p-values for the one-sided tests for treatment effect in a fixed-effects model, it is not that simple when using a mixed-effects model. The reason for this is that in a hierarchical setting, it becomes difficult to count the degrees of freedom which are required to set the distribution for the test statistic. If the variance of the fixed-effects component is known then one can easily obtain a p-value. However, the

variance is not known and is estimated using REML, hence there will be some degree of uncertainty surrounding the variance estimates. As a result, the cumulative distribution function of the test statistic will be unknown and thus a p-value cannot be calculated (157, 160). A number of approaches have been proposed to better approximate the degrees of freedom and thus obtain an approximate p-value (163, 164). However there is still much debate about how to best approximate the p-values and so there is no simple solution. Despite the ongoing debate, if a p-value is required, a sensible distributional assumption for the test statistic can be made based on the data and thus an approximate p-value obtained. More specifically in the context of IPD, due to the large amounts of data, one can treat the t-test statistic from the model as a z-statistic or z-score. Hence, a p-value from a one-sided test can then be estimated using a standard normal distribution which doesn't require the degrees of freedom to be specified. Thereafter, the one-sided z-statistics can be substituted into the splitting criterion to estimate the p-value of the interaction effect to aid the SIDES tree growing procedure. It is worth highlighting here that the p-values are not estimated to test hypotheses; rather they are utilized by the SIDES procedure to aid the exploratory search process. This therefore further justifies the distributional assumptions made for the test statistics in the fixed component of the mixed model. From now onwards, the extended SIDES method will be referred to as IPD-SIDES, which refers to three methods corresponding to models A, B and C above.

Now that splitting criteria have been defined for the IPD-IT and IPD-SIDES methods, what we now require is to determine which of the splitting criteria associated with models A, B and C provides reliable estimates of the splitting criterion score in order to detect interaction effects or identify subgroups with large treatment effects. The choice of model and thus the associated splitting criterion can be determined by performing a

simulation study considering different subgroup scenarios as well as varying degrees of between-study heterogeneity. The simulation study setup will now be described.

8.5 Simulation study setup

Data generation procedure

The simulation setup to evaluate the IPD-IT and IPD-SIDES method required that data be generated for several studies thus forming a pooled dataset. The same simulation setup described in Chapter 7 (section 7.2) was used to generate data for a single trial. Recall from Chapter 7 that the data generation process simply requires the main effects and interaction effects to be specified in order to generate the data. The same data generation procedure was applied several times to generate multiple single trial data that were then combined to form a pooled IPD dataset. However, in addition, this required some between-study variation, typically denoted by τ^2 in meta-analyses, to be introduced to represent the clustering present in the hierarchical data structure. This was done by randomly generating an overall mean β_0 for each trial in the pooled dataset from a normal distribution $\beta_0 \sim N(0, \tau^2)$ with a mean of zero and a between-trial variance of τ^2 , where τ^2 is to be specified. It is probably worth noting here that the ratio of the between-study variation τ^2 over the total variation provides another measure of heterogeneity commonly used in meta-analyses referred to as the intraclass correlation (ICC). Hence, the ICC measures how much of the total variance is explained by the clustering within each study. Values of the ICC typically range from zero to one where a value of zero suggests that the response values in one study are similar to those in other studies and a value of one suggests they are completely different.

Simulation study design

A full factorial simulation study design was used where four factors were varied; the sample size of each study in the pooled dataset, the interaction effect size for $T \cdot X_1$, the interaction effect size for $T \cdot X_2$ and the between-study variance. Each pooled dataset consisted of 5 studies with each study having a sample size of 200, 500 and 1000. For simplicity, the sample size of each study in the generated pooled dataset remained fixed. Furthermore, the simulated data assumed an equal proportion of individuals in each category of the treatment variable T , and the two covariates X_1 and X_2 .

Standardized interaction effect sizes of 0, 0.2 (small), 0.5 (medium), 0.8 (large) and 1.5 (very large) were considered for both the $T \cdot X_1$ and $T \cdot X_2$ interactions. Between-study variances (τ^2) of 0.1 (small) and 0.9 (large) relative to a residual within study variance of 1 were considered as they are within range of the typical between-study heterogeneity found in IPD meta-analyses (156, 165). These small and large between-study variances equate to ICC values of approximately 0.08 and 0.42 respectively in the simulated datasets. Varying τ^2 will enable us to investigate and contrast how the methods perform in varying degrees of between-study heterogeneity. In total, each permutation in the full-factorial design was simulated 1000 times and the results summarized. The simulations were performed using R software 3.0.1.

8.6 Simulation study results

IPD-IT Simulation Results

The results for the IPD-IT procedure using models A, B and C when there is small (0.1) between-study variation are presented in tables 8.1, 8.2 and 8.3 respectively. In this case as there is little between-study variation, we might expect all models to be ok. The results of the simulation study suggest that all three models perform similarly in the presence of small between-study variance. The small differences that are observed when comparing the results are quite likely due to simulation error. All three models

detect the correct tree more than 90% of the time when there are no interactions present. When only a single one-way interaction is present, all three models detect medium, large and very large interactions the majority of the time for all sample sizes ≥ 1000 and the models also do well in detecting a small one-way interaction when the sample size is 5000. When two one-way interactions are present, all three models detect large and very large interactions more than 97% of the time for all sample sizes ≥ 1000 . When both two-way interactions are of medium size, or when one of them is of medium size and the other is either large or very large, then all three models detect the correct tree the majority of the time provided the overall sample size is 2500 or more.

The results for the IPD-IT procedure using models A, B, and C when there is large (0.9) between-study variation are presented in tables 8.4, 8.5 and 8.6 respectively. All three models detect the correct tree a majority of the time when there are no interactions present. However, it can be observed that there is a slight decrease in the proportion of times the correct tree is detected when either a fixed-effect model (Model B) or mixed-effect model (Model C) is used i.e. slight increase of false positive rate. When there is a single one-way interaction present, the fixed-effect and mixed-models perform better than the null model (Model A) when a small interaction effect is present, in particular, they detect a small interaction a majority of the time if the sample size is 5000. All three models detect medium, large and very large single one-way interactions more than 90% of the time where the null model seems to perform a bit better than the fixed-effect and mixed-effect models. When there are two one-way interactions present, it is most noticeable that the fixed-effect and mixed-effect models in general perform far better than the null model. All three models detect the correct tree a majority of the time when the two-way interactions are either large or very large for all sample sizes ≥ 1000 . Moreover, the method detects the correct tree a majority of the

time when one of the interactions is of medium size and the other interaction is either medium, large or verge large for sample sizes greater than 2500.

IPD-SIDES Simulation Results

The results for the IPD-SIDES procedure using models A, B, and C when there is small (0.1) between-study variation are presented in tables 8.7, 8.8 and 8.9 respectively.

Similar to the IPD-IT method, all three models seem to perform rather similarly when there is small between-study variation. The method correctly detects the full dataset i.e. no subgroups, the majority of the time when no interactions are present. When there is a single one-way interaction present i.e. a single region with enhanced treatment effect, the IPD-SIDES method detects the correct subgroup more than 95% of the time for sample sizes ≥ 2500 for a medium sized interaction effect, and for all sample sizes ≥ 1000 for a large interaction effect size. When the one-way interaction is very large, there seems to be an unusual and rather large drop in the performance of the method. For example, it detects the correct subgroup approximately 98% of the time when the overall sample size is 1000 however it does not detect any correct subgroups at all when the sample size is 5000. This peculiarity will be investigated later. When there are two one-way interactions present i.e. two subgroups with enhanced treatment effect, the method detects the correct subgroups the majority of the time when the interactions are medium, large and very large for all sample sizes ≥ 1000 . Moreover, the method detects the correct subgroups a majority of the time when one of the interactions is small and the other interaction is very large for sample sizes ≥ 2500 .

The results for the IPD-SIDES procedure using models A, B and C when there is large (0.9) between-study variation are presented in tables 8.10, 8.11 and 8.12 respectively.

When no interactions are present, all three models detect the full dataset i.e. no

subgroups, the majority of the time however it is noticeable that the null model (Model A) performs better in this scenario. In general, it is most noticeable that the fixed-effect and the mixed-effect models perform far better than the null model when there is large between-study variance. For example, when there are two medium sized interactions and the sample size is 1000, the null model detects the correct subgroups around 70% of the time whereas the fixed-effect and mixed-effect models detect the correct subgroups around 85% of the time. Again, there is an unusual drop in performance of the method when there is a single very large one-way interaction present. Similar to the results of when the between-study variation was small, when there are two one-way interactions present, the method detects the correct subgroups the majority of the time when the interactions are medium, large and very large for all sample sizes ≥ 1000 .

8.7 Discussion

This chapter described the proposed extension of the IT and SIDES procedures to an IPD meta-analyses setting. The proposed extensions (IPD-IT and IPD-SIDES) were then evaluated using a simulation study to assess which model(s) (Model A - null model, Model B - fixed-effect model or Model C - mixed-effect model) provide the best estimate of the relevant test statistic to be used in the splitting criterion of both methods. The results of the simulation study for both methods suggest that when the between-study variation is small, the choice of model to estimate the splitting criterion does not matter i.e. it can be treated as a single dataset and the original method applied. On the other hand, when there is large between-study variation, both methods perform much better when either a fixed-effect or mixed-effect model is used to estimate the splitting criterion. The results of the simulation study therefore demonstrate that in the case of large between-study variation, the incorrect simpler model without trial effects does lead to a real loss in performance. In addition, the simulation results of the IPD-SIDES method highlighted an issue associated with detecting subgroups with larger effects.

The results of the simulation study are reflective of the approaches currently undertaken to analyse data that have a hierarchical structure. Typically, the data are firstly assessed for the degree of between-study variation using a measure such as the ICC. If the ICC is small and thus negligible, then standard statistical analyses are implemented, otherwise, alternative models such as fixed or mixed models are applied to account for the clustering. Thus, in the same manner, it is recommended here that the between-study variation is assessed beforehand to help determine which model to use to estimate the splitting criterion for both the IPD-IT and IPD-SIDES methods when applying them to individual patient data.

This chapter has thus developed and evaluated two statistical methods, namely IPD-IT and IPD-SIDES, for performing subgroup analyses in an IPD meta-analyses setting. Although the IPD-SIDES method performs quite well in most of the scenarios considered, the performance of the method somehow deteriorates when detecting subgroups with a very large effect. It is therefore worth investigating why this issue occurs and possibly find a solution to the problem. The next chapter will therefore thoroughly investigate the issue with the IPD-SIDES method with the aim of trying to rectify the problem and thus improving the method.

Table 8.1 – Simulation results for the IPD-IT method for Model A (Null model) when there is small (0.1) between-study variation. Results display % of correctly identified final trees.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	94.50	18.70	86.10	92.80	94.10
	2500	93.90	51.20	90.70	91.90	93.90
	5000	93.90	79.30	93.20	93.50	92.70
Small=0.2	1000	22.70	0.70	4.10	5.10	4.10
	2500	54.30	9.10	21.40	20.70	20.80
	5000	80.00	35.90	48.70	51.50	55.70
Medium=0.5	1000	87.10	3.70	47.90	66.90	65.10
	2500	91.60	20.80	96.50	98.40	98.40
	5000	93.00	52.90	100.00	100.00	100.00
Large=0.8	1000	92.50	4.60	64.60	97.80	98.80
	2500	93.80	21.10	98.50	100.00	100.00
	5000	92.00	53.50	100.00	100.00	100.00
V. Large=1.5	1000	93.20	5.20	64.80	98.70	100.00
	2500	92.00	19.90	98.20	100.00	100.00
	5000	93.50	53.60	100.00	100.00	100.00

Table 8.2 – Simulation results for the IPD-IT method for Model B (Fixed-effect model) when there is small (0.1) between-study variation. Results display % of correctly identified final trees.

T^*X_1 Standardized interaction effect size	N	T^*X_2 Standardized interaction effect size				
		None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	93.90	24.60	86.20	88.70	91.30
	2500	93.30	54.00	90.80	90.80	92.30
	5000	92.20	78.40	91.80	92.80	92.00
Small=0.2	1000	24.00	0.80	6.60	6.50	7.90
	2500	56.40	8.90	24.50	20.60	21.80
	5000	82.20	35.00	54.30	57.10	57.70
Medium=0.5	1000	85.90	6.20	52.80	69.30	66.60
	2500	89.50	24.30	97.60	98.90	98.70
	5000	91.60	56.00	100.00	100.00	100.00
Large=0.8	1000	89.40	5.90	68.60	97.50	98.70
	2500	91.70	23.20	98.60	100.00	100.00
	5000	93.20	52.50	100.00	100.00	100.00
V. Large=1.5	1000	89.90	4.80	68.00	99.00	100.00
	2500	92.10	20.60	98.90	100.00	100.00
	5000	92.30	54.80	100.00	100.00	100.00

Table 8.3 – Simulation results for the IPD-IT method for Model C (Mixed-effect model) when there is small (0.1) between-study variation. Results display % of correctly identified final trees.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	91.80	22.30	85.00	89.70	91.50
	2500	92.60	55.10	88.90	90.90	92.70
	5000	93.20	80.40	91.60	91.60	92.40
Small=0.2	1000	25.60	1.00	4.60	6.60	5.90
	2500	51.10	9.40	21.90	23.10	24.20
	5000	82.00	36.00	57.50	55.40	56.70
Medium=0.5	1000	83.20	6.60	50.40	66.00	69.10
	2500	90.40	21.40	98.30	98.70	98.80
	5000	90.60	55.90	100.00	100.00	100.00
Large=0.8	1000	90.30	6.60	70.70	97.50	98.70
	2500	91.50	20.70	98.60	100.00	100.00
	5000	91.20	55.40	100.00	100.00	100.00
V. Large=1.5	1000	90.70	6.00	69.60	98.50	100.00
	2500	91.60	21.90	98.90	100.00	100.00
	5000	93.20	55.70	100.00	100.00	100.00

Table 8.4 – Simulation results for the IPD-IT method for Model A (Null model) when there is large (0.9) between-study variation. Results display % of correctly identified final trees.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	98.40	10.10	81.60	97.00	97.60
	2500	98.20	39.50	97.30	97.10	98.40
	5000	98.70	70.10	96.90	97.20	98.40
Small=0.2	1000	11.00	0.10	2.20	1.30	1.50
	2500	33.30	2.90	9.20	9.70	9.60
	5000	72.70	18.00	28.90	32.40	29.90
Medium=0.5	1000	82.10	1.80	28.90	45.00	43.00
	2500	96.30	10.70	90.70	94.90	94.20
	5000	97.20	30.60	99.90	100.00	99.90
Large=0.8	1000	97.50	1.20	44.60	90.40	96.10
	2500	98.20	9.40	94.30	100.00	100.00
	5000	97.20	34.30	99.90	100.00	100.00
V. Large=1.5	1000	96.90	1.50	45.50	94.20	100.00
	2500	96.80	8.40	94.00	100.00	100.00
	5000	97.60	33.00	100.00	100.00	100.00

Table 8.5 – Simulation results for the IPD-IT method for Model B (Fixed-effect model) when there is large (0.9) between-study variation. Results display % of correctly identified final trees.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	93.90	22.40	85.50	90.20	90.60
	2500	93.20	53.30	90.90	91.30	90.80
	5000	93.00	80.60	90.00	91.50	91.80
Small=0.2	1000	22.50	1.40	5.80	5.40	5.70
	2500	55.80	9.60	19.90	21.70	23.30
	5000	80.60	37.60	54.20	53.70	55.40
Medium=0.5	1000	87.40	4.00	52.00	67.50	65.90
	2500	90.30	21.90	98.00	99.30	98.80
	5000	91.40	56.60	100.00	100.00	100.00
Large=0.8	1000	90.00	6.20	66.40	98.60	99.00
	2500	90.90	24.50	98.60	100.00	100.00
	5000	92.90	55.40	100.00	100.00	100.00
V. Large=1.5	1000	90.90	6.50	66.30	98.40	100.00
	2500	91.30	24.90	98.10	100.00	100.00
	5000	92.40	51.70	100.00	100.00	100.00

Table 8.6 – Simulation results for the IPD-IT method for Model C (Mixed-effect model) when there is large (0.9) between-study variation. Results display % of correctly identified final trees.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	92.60	23.30	87.40	91.00	90.40
	2500	93.90	56.30	89.20	91.90	90.80
	5000	94.10	81.20	90.80	91.20	90.60
Small=0.2	1000	22.70	0.60	6.40	6.30	6.00
	2500	55.00	10.20	23.20	22.80	23.20
	5000	82.00	37.40	55.50	55.30	55.80
Medium=0.5	1000	86.70	5.90	51.20	67.10	67.50
	2500	90.30	21.40	97.40	98.70	98.40
	5000	92.20	55.10	100.00	100.00	100.00
Large=0.8	1000	90.40	5.10	67.30	97.90	99.40
	2500	90.80	21.00	98.30	100.00	100.00
	5000	91.40	53.60	100.00	100.00	100.00
V. Large=1.5	1000	89.70	5.30	69.60	98.80	100.00
	2500	91.60	21.40	99.20	100.00	100.00
	5000	92.70	52.60	100.00	100.00	100.00

Table 8.7 – Simulation results for the IPD-SIDES method for Model A (Null model) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	89.50	17.90	78.60	97.10	98.50
	2500	90.30	44.20	96.30	98.00	7.20
	5000	91.70	70.40	97.50	88.00	0.00
Small=0.2	1000	19.40	15.00	26.70	21.30	13.60
	2500	46.40	43.10	54.30	45.50	97.40
	5000	73.00	74.50	78.30	76.60	100.00
Medium=0.5	1000	82.60	23.40	83.70	89.30	82.60
	2500	96.40	49.10	100.00	99.70	100.00
	5000	98.30	79.00	100.00	100.00	100.00
Large=0.8	1000	97.50	20.70	90.20	100.00	99.50
	2500	98.00	42.50	99.90	100.00	100.00
	5000	93.60	78.10	100.00	100.00	100.00
V. Large=1.5	1000	98.70	14.10	82.20	99.80	100.00
	2500	55.60	68.90	100.00	100.00	100.00
	5000	47.90	89.80	100.00	100.00	100.00

Table 8.8 – Simulation results for the IPD-SIDES method for Model B (Fixed-effect model) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	88.50	19.10	77.40	95.90	98.30
	2500	90.40	46.50	95.80	97.40	4.70
	5000	90.10	70.60	97.60	82.30	0.00
Small=0.2	1000	23.90	16.70	30.30	24.50	19.00
	2500	45.70	46.30	56.70	46.90	99.30
	5000	71.30	78.20	81.40	83.00	100.00
Medium=0.5	1000	79.60	30.70	85.30	91.50	84.70
	2500	97.20	53.20	99.90	99.70	100.00
	5000	97.70	80.70	100.00	100.00	100.00
Large=0.8	1000	96.00	23.60	91.00	99.80	100.00
	2500	98.00	45.80	99.70	100.00	100.00
	5000	89.10	83.90	100.00	100.00	100.00
V. Large=1.5	1000	98.90	19.00	86.20	99.90	100.00
	2500	52.10	75.60	100.00	100.00	100.00
	5000	50.10	90.60	100.00	100.00	100.00

Table 8.9 – Simulation results for the IPD-SIDES method for Model C (Mixed-effect model) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups.

T^*X_1 Standardized interaction effect size	N	T^*X_2 Standardized interaction effect size				
		None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	88.70	21.00	74.50	95.90	97.10
	2500	88.10	43.80	95.00	97.50	2.50
	5000	89.10	70.70	97.00	78.30	0.00
Small=0.2	1000	20.80	16.60	32.30	26.10	18.90
	2500	43.70	43.30	55.10	48.20	99.40
	5000	73.40	79.00	79.80	83.70	100.00
Medium=0.5	1000	78.00	29.70	86.60	91.00	85.30
	2500	95.50	55.40	99.70	99.90	100.00
	5000	97.40	77.50	100.00	100.00	100.00
Large=0.8	1000	95.10	22.80	90.90	99.70	99.70
	2500	97.30	47.00	99.80	100.00	100.00
	5000	87.10	82.90	100.00	100.00	100.00
V. Large=1.5	1000	97.00	18.00	83.20	99.90	100.00
	2500	52.10	77.40	100.00	100.00	100.00
	5000	50.70	90.70	100.00	100.00	100.00

Table 8.10 – Simulation results for the IPD-SIDES method for Model A (Null model) when there is large (0.9) between-study variation. Results display % of correctly identified final subgroups.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	95.40	10.60	68.50	96.90	99.90
	2500	95.90	28.10	96.30	99.40	68.70
	5000	96.00	58.50	99.00	98.60	7.30
Small=0.2	1000	14.90	8.20	21.80	12.20	7.00
	2500	32.20	30.00	38.60	27.90	51.40
	5000	61.00	60.60	62.60	59.50	98.10
Medium=0.5	1000	70.00	17.30	70.80	79.00	67.4
	2500	98.00	36.90	99.20	99.20	98.50
	5000	99.80	61.10	100.00	100.00	100.00
Large=0.8	1000	98.00	10.90	75.60	99.50	99.10
	2500	99.30	27.40	99.50	100.00	100.00
	5000	99.30	55.90	100.00	100.00	100.00
V. Large=1.5	1000	99.90	6.70	67.80	98.40	100.00
	2500	82.40	37.90	98.70	100.00	100.00
	5000	53.20	79.60	100.00	100.00	100.00

Table 8.11 – Simulation results for the IPD-SIDES method for Model B (Fixed-effect model) when there is large (0.9) between-study variation. Results display % of correctly identified final subgroups.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	90.50	22.20	79.80	96.20	98.00
	2500	87.60	44.30	96.70	97.70	4.20
	5000	88.60	70.60	97.60	82.50	0.00
Small=0.2	1000	24.00	16.90	30.60	25.70	14.40
	2500	45.00	45.90	56.40	47.20	98.70
	5000	71.10	77.90	79.50	85.10	100.00
Medium=0.5	1000	81.50	30.10	85.00	90.90	85.40
	2500	95.20	52.50	99.90	99.80	100.00
	5000	98.10	82.60	100.00	100.00	100.00
Large=0.8	1000	95.70	21.20	90.80	99.60	99.40
	2500	98.60	48.50	99.50	100.00	100.00
	5000	90.20	82.40	100.00	100.00	100.00
V. Large=1.5	1000	98.80	14.80	84.80	99.40	100.00
	2500	50.50	70.50	100.00	100.00	100.00
	5000	50.90	92.30	100.00	100.00	100.00

Table 8.12 – Simulation results for the IPD-SIDES method for Model C (Mixed-effect model) when there is large (0.9) between-study variation. Results display % of correctly identified final subgroups.

		T^*X_2 Standardized interaction effect size				
T^*X_1 Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	88.60	20.30	77.80	96.40	96.00
	2500	89.60	44.70	96.20	96.60	2.00
	5000	89.50	72.00	97.40	76.00	0.00
Small=0.2	1000	20.80	18.30	31.00	25.10	19.10
	2500	44.70	44.40	54.80	46.50	99.60
	5000	72.70	75.40	80.10	84.60	100.00
Medium=0.5	1000	80.80	31.20	85.60	92.00	85.60
	2500	95.40	54.00	100.00	99.60	100.00
	5000	97.70	80.00	100.00	100.00	100.00
Large=0.8	1000	95.90	24.00	90.50	99.80	99.70
	2500	98.20	48.40	100.00	100.00	100.00
	5000	87.60	81.10	100.00	100.00	100.00
V. Large=1.5	1000	97.00	17.80	85.20	99.90	100.00
	2500	48.40	75.40	100.00	100.00	100.00
	5000	49.30	90.60	100.00	100.00	100.00

Chapter 9

Further development of the IPD-SIDES method

9.1 Introduction

The previous chapter developed and evaluated the IPD-IT and IPD-SIDES methods for performing subgroup analyses in an IPD meta-analyses framework. Both of the methods seemed to perform rather well, especially when there is large between-study variance present. However, it was observed from the IPD-SIDES simulation results that there was a rather unusual decrease in the performance of the method when there is a single very large one-way interaction present. The consideration of much larger samples in the simulation study in the previous chapter may have magnified a problem that was not noticeable in the single trial scenario simulation presented in Chapter 7.

Since the IPD-SIDES method works quite well in most of the scenarios considered in the simulation study, it is worth investigating how and why the methods performance is affected in the hope that a solution can be found. This chapter will therefore thoroughly investigate the issue to try and better understand why the drop in

performance occurs. Moreover, possible solutions to the problem will be sought and evaluated by re-running the simulation study conducted in the previous chapter.

9.2 Investigating the drop in performance of the IPD-SIDES method

The first step to better determine why the issue occurs is to observe what happens when applying the method to a single simulated dataset. The single dataset was generated ($N=5000$) with a very large single one-way interaction and the IPD-SIDES method was applied. Table 9.1 displays the final subgroups detected when applying the method. The method should only detect one subgroup (i.e. the first row only in Table 9.1), however it goes on to split further and detect another subgroup. One of the possible reasons for this happening is because there is no restriction placed on the computed differential effect p-value i.e. the splitting criterion (final column in Table 9.1). To recall how the method works, first the method evaluates the differential effect splitting criterion for all possible splits and retains the best single split i.e. split with the smallest p-value, for each covariate. These splits are then ordered in terms of the differential effect p-value from smallest to largest. Thereafter, only the top M splits with the largest differential effect are considered where M is pre-specified by the user. The ordered splits, though they may not be significant, are explored individually where only the subgroup with the more enhanced treatment effect is retained from each split provided it meets the continuation criterion. Hence, the differential effect splitting criterion is merely used to order the splits regardless of their significance. Therefore, it is possible for the method to detect a spurious subgroup resulting from a split with a non-significant differential effect. Thus, one possible solution to the problem would be to impose a restriction on the computed differential effect p-values such that the procedure only considers splits where the p-value is within a certain threshold e.g. 0.05 or 0.10. If such a restriction is applied to the example in Table 9.1, then only one subgroup would have been identified. It is worth mentioning here that the splitting

criterion is used to aid the search process of the method to identify candidate subgroups. Therefore, the restriction placed on the splitting criterion p-value need not be that strict ($p \leq 0.05$), and instead a less stringent restriction can be imposed ($p \leq 0.10$) so that the method has the flexibility to identify subgroups that might be plausible.

Another reason why the procedure goes on to detect additional subgroups is because of the continuation criterion used by the method. Recall from chapter 6 that the method only keeps a split provided it satisfies the continuation criterion $p_c \leq \gamma \cdot p_p$ i.e. that the one-sided treatment effect p-value of the newly formed child node must be less than or equal to the complexity parameter value multiplied by the one-sided treatment effect p-value of the parent node from which it came. As we are dealing with large sample sizes, the treatment effect estimates in the simulated datasets will have more precision i.e. smaller standard error. Therefore, if the effect size is large, then dividing by a very small standard error will produce an extremely small one-sided p-value for the treatment effect. Often the p-values are so small that computational limitations means that they are calculated or stored as zeros. Therefore, the first subgroup selected in Table 9.1 (first row) has a treatment effect of 0.808 and has a one-sided p-value of zero. Subsequently, the method then goes on to select a second subgroup (second row in Table 9.1) because the one-sided p-value is equal to the p-value of the parent node from which it came; thus satisfying the continuation criterion. Thus, a solution to this problem would be to change the continuation criterion so that the one-sided p-value of the child node should be strictly less than the one-sided p-value of the parent node i.e. $p_c < \gamma \cdot p_p$. However, if we think about the aim of the method, the aim is to identify several candidate subgroups with enhanced treatment effect. Therefore, changing the inequality to being strictly less than will certainly disregard several candidate subgroups that are similar to the parent node from which they were formed.

Thus, changing the inequality might not be a suitable solution with regards to the objective of the method.

From observing the output from the single simulated dataset, another issue with the method becomes apparent. Notice how the second subgroup identified by the method has a treatment effect of 0.843. This treatment effect is actually not that different from the treatment effect in the disregarded subgroup (0.773). The reason the method considers this split is because of the differential effect splitting criterion used by the method. Recall that this splitting criterion is of the form

$$p = 2 \cdot \left[1 - \Phi \left(\frac{|Z_{E1} - Z_{E2}|}{\sqrt{2}} \right) \right] \quad (9.1)$$

where Z_{E1} and Z_{E2} are the one-sided test statistics from the tests computed in child nodes 1 and 2 respectively. As explained earlier, the splitting criterion is only used to order all of the splits in terms of the differential effect p-value. Thereafter, the best M splits are considered from which the subgroups with larger treatment effect are retained (provided they meet the continuation criterion). Though the authors propose this splitting criterion to perform the differential effect search, it is probably not the most appropriate to directly evaluate the differential effect. As a result, it is quite possible that larger differential effects will go unnoticed. The reason why it is not the most appropriate splitting criterion is because it can give a significant p-value when the treatment effects in the two child nodes are similar and the standard errors are sufficiently different. In this case the difference in standardized Z statistics in (9.1) may be large even if the difference in (non-standardized) effect sizes is small. To consider an example, say a split is formed where both child nodes have a treatment effect of 4.0, however the SE in the left child node is 0.2 and the SE in the right child node is 0.9.

Computation of the one-sided test statistics $Z_{E1} = \frac{4.0}{0.2} = 20$ and $Z_{E2} = \frac{4.0}{0.9} = 4.4$ would

suggest that there is a big differential effect between the two groups and thus the splitting criterion defined by equation (9.1) would indicate that the differential effect is highly significant. Now if the same split was evaluated using a regression model with the inclusion of an interaction effect, then the test statistic for the interaction effect would be very small thus being indicative of there being no differential effect present. What this means is that the current splitting criterion is more likely to select splits where there is a larger subgroup with more precision i.e. smaller SE, compared to a smaller subgroup with less precision i.e. larger SE, regardless of whether the treatment effects are the same or not. The current splitting criterion is therefore not an appropriate approach for performing the differential effect search for the objective we wish to use it for and so another alternative criterion is required. As mentioned earlier on in this thesis, the most appropriate method of directly evaluating a differential subgroup effect is to use a statistical test for interaction. Therefore, one approach to get the method to do what we require it to do is to define a new splitting criterion that uses the interaction effect estimate test statistic to obtain a differential effect p-value. We can define the new splitting criterion as follows

$$p = 2 \cdot [1 - \Phi(Z_{int})] \quad (9.2)$$

where Z_{int} is the two-sided hypothesis test statistic computed for the interaction effect estimate that is obtained from fitting an appropriate regression model (linear, fixed or mixed model) with the inclusion of an interaction term and where $\Phi(Z_{int})$ is the cumulative distribution function of the standard normal distribution.

A closer inspection of the method coding provided by the authors identified yet another important issue with the method. It was found that having ordered the splits using the original splitting criterion; the method selects the subgroups with a larger one-sided test statistic i.e. smaller p-value rather than selecting the subgroups with the

largest treatment effect. Thus the method aims to find subgroups with the smallest p-value rather than largest treatment effect, which is not what we want the method to do. Therefore, the code provided by the authors doesn't actually do what we require it to do. In order for the method to do what we require, the coding can be changed such that the method selects the subgroup with the largest treatment effect instead of selecting the subgroup with the largest one-sided test statistic. In this way, the objective of the method changes to identifying subgroups with enhanced treatment effect rather than identifying subgroups with smallest p-value. From now onwards, this will be referred to as the modified IPD-SIDES method.

9.3 Simulation Study

The previous section investigated three issues with using the IPD-SIDES method to do what we want; firstly an issue highlighted in the previous chapter regarding the method's performance when considering larger effect sizes, secondly an issue recognized in the previous section regarding the differential effect splitting criterion and thirdly an issue with the method coding in the way it selects subgroups. A possible solution for each of these issues was described in the previous section. In this section, we report the evaluation of the modified method using simulation studies. Two simulation studies were performed. Simulation study 1 evaluated the performance of the IPD-SIDES method proposed in the previous chapter with a restriction imposed on the p-values ($p \leq 0.10$) computed using the differential effect search splitting criterion. Simulation study 2 evaluated the performance of the modified IPD-SIDES method that uses a new splitting criterion defined by equation (9.2) and also selects the subgroup with the largest treatment effect (instead of the selecting the subgroup with smallest p-value as done so by the original method). Moreover, a restriction was also placed on the p-values ($p \leq 0.10$) obtained from the interaction. Both of these simulation studies were set up in exactly the same way as described in the previous chapter. As there was

not much difference in the simulation study results of the method in the previous chapter when a fixed-effect or mixed-effect model was used, the simulation study in this chapter only considered the splitting criterion that utilizes a mixed-effect model. Moreover, simulation study 1 and 2 both investigated the performance of the method in the presence of small (0.1) and large (0.9) between-study variance.

9.4 Simulation Study 1 results

The results of simulation study 1 are presented in tables 9.2 (small (0.1) between-study variation) and 9.3 (large (0.9) between-study variation) respectively. The results were quite similar in both of the results tables. When there are no subgroups to detect, the method detects the correct subgroup (full dataset) around 95% of the time. This is an improvement compared to the results in the previous chapter which observed a drop in the performance of the method when using either a fixed or mixed model when compared to the null model. The method detects the correct subgroup a majority of the time when there is a single medium, large or very large one-way interaction present. It can be observed from the results that imposing a restriction on the differential effect p-value has resolved the issue highlighted in the previous chapter and the method now detects very large effects more than 90% of the time. When there are two one-way interactions present that are medium, large or very large, then the method detects the correct subgroups a majority of the time. When one of the interactions is small and the other interaction is medium, large or very large, then the method detects the correct subgroups a majority of the time provided the overall sample size is ≥ 5000 .

9.5 Simulation Study 2 results

The results of simulation study 2 are presented in tables 9.4 (small between-study variation) and 9.5 (large between-study variation) respectively. The results in both

tables were quite similar and can be interpreted in the same way. When no interactions are present, the modified IPD-SIDES method detects the full dataset around 95% of the time. When there is a single medium, large or very large one-way interaction present, the method detects the correct subgroups the majority of the time. The modified method did not display any issues when detecting very large single interaction effects as was observed in the previous chapter. When two medium, large or very large one-way interactions are present, the method detects the correct subgroups with large treatment effect the majority of the time. When one of the interactions is small and the other interaction is medium, large or very large, then the method detects the correct subgroups a majority of the time provided the overall sample size is ≥ 5000 . In general, the results of the modified IPD-SIDES method are very promising in that it detects the correct subgroups with large treatment effect a majority of the time.

9.6 Discussion

This chapter set out to investigate an issue identified with the IPD-SIDES method proposed in the previous chapter. A restriction was imposed on the splitting criterion p-value as a solution to the problem and the results of the simulation study suggests that this resolves the issue. The method also does something different to what we require it to do i.e. it aims to identify subgroups with small p-values. Hence, the method was further developed to better suit the objective of the work in this thesis and thus the modified IPD-SIDES method was proposed and evaluated. The results of the new modified method were very promising in detecting subgroups with large or enhanced treatment effect.

This chapter proposes the modified IPD-SIDES method as a useful tool for performing subgroup analyses in and IPD meta-analyses setting where the aim is to identify several candidate subgroups with large treatment effect. Having developed and

evaluated the modified IPD-SIDES method in a number of simulated settings, it is now of interest to apply this method to some real data. The next chapter will therefore apply both the IPD-IT and the modified IPD-SIDES methods to real individual patient data from several low back pain trials. Moreover, the IPD-SIDES method proposed in the previous chapter (using the original splitting criterion) and the modified IPD-SIDES method proposed in this chapter (using the restricted interaction test splitting criterion) will both be applied to the real data and compared to better demonstrate the difference between the two methods in terms of the subgroups they identify.

Table 9.1 – Simulation output for when there is a $t \times v_1$ interaction effect of 1.5 and total sample size $N=5000$.

	Selected subgroup				Disregarded subgroup					
Subgroup	n1	Trt1	SE Trt1	Trt1 p-value	n0	Trt0	SE Trt0	Trt0 p-value	Differential effect	Split criterion p-value
x1 > -0.5	2500	0.808	0.041	0.000	2500	-0.760	0.040	1.000	1.571	0.000
x1 > -0.5 & x2 > -0.5	1250	0.843	0.058	0.000	1250	0.773	0.058	0.000	0.069	0.425

Table 9.2 – Simulation results for the IPD-SIDES method with a restriction on the differential effect p-value ($p \leq 0.10$) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups.

		$T \times X_2$ Standardized interaction effect size				
$T \times X_1$ Standardized interaction effect size	N	None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	96.20	19.30	83.40	97.60	97.90
	2500	94.70	48.30	96.80	97.90	93.00
	5000	96.70	78.30	98.40	94.80	93.10
Small=0.2	1000	17.70	4.30	17.70	17.80	17.20
	2500	49.20	28.00	47.90	55.30	61.50
	5000	77.30	64.90	78.80	82.80	86.10
Medium=0.5	1000	84.10	16.40	84.50	86.20	85.80
	2500	97.50	46.20	99.70	100.00	100.00
	5000	98.10	78.00	100.00	100.00	100.00
Large=0.8	1000	98.00	17.80	90.90	99.90	99.90
	2500	98.30	44.60	100.00	100.00	100.00
	5000	96.10	82.50	100.00	100.00	100.00
V. Large=1.5	1000	98.10	18.60	91.30	99.80	100.00
	2500	96.30	58.20	100.00	100.00	100.00
	5000	96.70	86.10	100.00	100.00	100.00

Table 9.3 – Simulation results for the IPD-SIDES method with a restriction on the differential effect p-value ($p \leq 0.10$) when there is small (0.9) between-study variation. Results display % of correctly identified final subgroups.

T^*X_1 Standardized interaction effect size	N	T^*X_2 Standardized interaction effect size				
		None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	96.30	18.60	86.60	97.90	97.20
	2500	96.30	46.90	98.20	98.50	94.50
	5000	96.20	76.30	98.20	95.90	94.20
Small=0.2	1000	18.60	4.90	19.70	18.00	17.60
	2500	47.40	29.10	47.20	55.30	61.10
	5000	77.90	66.20	80.10	86.50	86.90
Medium=0.5	1000	84.90	17.40	85.30	92.60	85.30
	2500	98.00	48.40	99.70	100.00	100.00
	5000	98.20	85.80	100.00	100.00	100.00
Large=0.8	1000	96.70	22.50	90.60	99.90	99.70
	2500	97.10	48.30	100.00	100.00	100.00
	5000	95.90	83.40	100.00	100.00	100.00
V. Large=1.5	1000	98.30	15.50	91.50	100.00	100.00
	2500	95.60	51.30	99.90	100.00	100.00
	5000	95.60	83.30	100.00	100.00	100.00

Table 9.4 – Simulation results for the modified IPD-SIDES with a restriction on the differential effect p-value ($p \leq 0.10$) when there is small (0.1) between-study variation. Results display % of correctly identified final subgroups.

T^*X_1 Standardized interaction effect size	N	T^*X_2 Standardized interaction effect size				
		None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	96.80	20.50	83.40	97.50	98.40
	2500	95.60	46.50	97.60	97.90	95.10
	5000	95.60	72.80	98.40	95.60	94.40
Small=0.2	1000	17.20	6.90	16.50	17.10	16.10
	2500	49.00	38.80	50.80	44.80	61.60
	5000	77.80	75.90	83.40	80.70	87.10
Medium=0.5	1000	85.10	16.90	77.00	91.60	83.20
	2500	97.90	54.60	100.00	100.00	100.00
	5000	97.90	80.20	100.00	100.00	100.00
Large=0.8	1000	97.70	16.70	83.70	99.80	99.50
	2500	98.10	47.40	99.90	100.00	100.00
	5000	96.10	81.80	100.00	100.00	100.00
V. Large=1.5	1000	98.30	17.20	85.40	99.50	100.00
	2500	95.60	53.00	100.00	100.00	100.00
	5000	96.40	82.50	100.00	100.00	100.00

Table 9.5 – Simulation results for the modified IPD-SIDES with a restriction on the differential effect p-value ($p \leq 0.10$) when there is small (0.9) between-study variation. Results display % of correctly identified final subgroups.

T^*X_1 Standardized interaction effect size	N	T^*X_2 Standardized interaction effect size				
		None=0	Small=0.2	Medium=0.5	Large=0.8	V. Large=1.5
None=0	1000	96.60	21.30	86.50	98.20	96.90
	2500	96.60	50.00	97.90	98.10	94.10
	5000	95.90	78.30	98.60	94.50	94.60
Small=0.2	1000	19.30	4.90	18.90	19.80	16.60
	2500	48.40	35.90	46.40	48.20	57.70
	5000	74.30	63.40	85.60	82.20	84.90
Medium=0.5	1000	83.40	21.90	84.30	90.90	83.40
	2500	98.60	44.60	100.00	99.70	99.90
	5000	98.20	79.10	100.00	100.00	100.00
Large=0.8	1000	97.90	28.10	83.60	99.70	99.80
	2500	98.40	43.80	99.80	100.00	100.00
	5000	96.80	87.00	100.00	100.00	100.00
V. Large=1.5	1000	98.50	15.00	85.20	99.40	100.00
	2500	96.40	54.40	100.00	100.00	100.00
	5000	96.70	82.70	100.00	100.00	100.00

PART III

Application and Summary

Chapter 10

Application of IPD-IT and IPD-SIDES to real data

10.1 Introduction

The previous two chapters (Chapters 8 and 9) described and implemented a simulation study to evaluate the performance of the proposed IPD-IT, unmodified IPD-SIDES and the modified IPD-SIDES methods. The results suggest that the IPD-IT and modified IPD-SIDES methods are promising approaches for performing subgroup analyses or subgroup identification in an IPD meta-analysis setting, particularly in detecting larger interaction effects. The choice of which model to use (linear regression, fixed model or mixed model) to estimate the splitting criterion depends on the degree of between-study variability present in the dataset. If the between-study variability is small and negligible then it does not matter which model is applied. However, when there is considerable between-study variation present then it is recommended that either a fixed-effect or mixed-effect model be used to estimate the splitting criterion.

Having assessed the proposed extensions in a number of simulation settings, it is now of interest to apply these methods to some real data. Therefore, this chapter will present the results of the application of the proposed methods to a pooled dataset of data from a number of low back pain (LBP) acupuncture trials. Initially the data from the individual trials contained in the pooled dataset will be described. The statistical analysis methods applied to these data will then be detailed. Finally, the results from the analyses will be presented and discussed. Both the unmodified IPD-SIDES and modified IPD-SIDES methods will be applied to the real data in this chapter to better visualize the difference between the two methods in the subgroups they identify.

10.2 Description of the pooled dataset

The pooled dataset consists of data from 4 acupuncture trials with short-term follow-up (8-13 weeks). The names of these trials will not be disclosed here and will be referred to as Trial 1, Trial 2, Trial 3 and Trial 4 respectively. In total, the pooled dataset consists of 4540 individuals; 298 from Trial 1, 1162 from Trial 2, 2841 from Trial 3 and 239 from trial 4. Though the data for each trial contained several demographic and baseline covariates, only three covariates were common to all four trials. All of the trials collected age and gender as demographic variables. In addition they also collected SF-36 data at baseline and short-term follow-up. The SF-36 is a questionnaire that measures health related quality of life (32). The items in the SF-36 questionnaire are used to create two aggregated summary measures; a physical component score (PCS) and a mental health component score (MCS). These are measured on a scale of 0-100 where a lower score indicates poorer physical or mental functioning.

10.3 Methods

Data formatting

Prior to performing any analyses, it is important to specify an appropriate form for the continuous dependent variable to be used in the statistical analyses. There are a number of ways in which continuous outcome data can be used in analyses. One approach is to use the raw outcome data as the dependent variable and adjust for the baseline scores. Other common approaches include using change from baseline or percentage change from baseline as the dependent variable. The change from baseline approach from a clinical perspective is more appealing due to the ease of interpretability. However, an issue with this measure is that it has floor and ceiling effects. For example a patient with a high RMDQ score at baseline has more room for improvement in RMDQ score at post-treatment follow-up than a patient with a low score at baseline. For this reason, it might be more sensible to use the percentage change from baseline as a better measure of change. However, clinicians may find it more difficult to interpret a percentage change from baseline score compared to a change from baseline score. Furthermore, it has been shown that using the percentage change from baseline is statistically inefficient in terms of power when compared to using the change from baseline score (166). Thus, having described the limitations of these approaches, selecting which approach to take can be a difficult task. An obvious place to start is by observing the distribution of each of the different forms that the dependent variable can take since the models fitted by the tree methods to estimate the spitting criteria require the data be normally distributed. The outcomes of interest in the pooled dataset were the MCS and PCS scores from the SF-36 questionnaire measured at baseline and at around 8-13 weeks (short-term) follow-up. The different distributions for both the short-term MCS and PCS follow-up scores were compared i.e. the raw scores, the change from baseline scores and the percentage change from

baseline scores. It was observed that the distribution of the follow-up raw scores, in particular the MCS, were skewed. The percentage change from baseline MCS and PCS scores were even more skewed. However, both the change from baseline MCS and PCS scores were normally distributed. For this reason, it was decided to use the change from baseline MCS and PCS short-term follow-up scores as the dependent variables in the analyses. The change from baseline scores were computed by simply subtracting the baseline score from the short-term follow-up score, where a positive change score would indicate improvement.

Both of the tree methods are designed to compare two arms e.g. intervention vs. control, therefore a binary treatment variable was created. Here, those who received GP care, usual care and sham acupuncture were classed as being in the control arm and those receiving acupuncture were classed as being in the intervention arm. The application of the tree methods requires that the dataset is complete. Therefore, 695 individuals in the dataset with missing data were removed prior to performing the analyses. Though there are methods for dealing with missing data e.g. imputation and surrogate splits (used by the CART procedure) (130), these were not considered here as the objective was to simply demonstrate the application of the extended methods. Finally, the MCS and PCS baseline scores were rounded to one decimal place and the age covariate was rounded to a whole number to reduce the computational time searching all potential splits. Just to give an example, the original baseline MCS score has 3,432 unique split points that will need to be searched by both methods. Rounding the values of this covariate to one decimal place vastly reduces the number of unique split points to be searched to 512.

Assessing risk of bias of main trial

The overall risk of bias in the main trial of each of the included trials was carried out using the Cochrane Collaboration risk of bias tool (167). The criteria used for the assessment were: 1) method of randomization, 2) allocation concealment, 3) blinding, 4) incomplete outcome data, 5) selective reporting, 6) similarity of groups at baseline, 7) sample size calculation, and 8) intention to treat analysis.

Statistical analyses

Prior to performing the main analyses, the demographic data, baseline data and outcome data were summarized for the pooled dataset. These summaries did allow for the clustering of the data. For continuous variables, the mean and standard deviation, median, range and frequency of missing data are presented. For categorical data, the frequency and percentage are presented for each category which also includes the frequency and percentage for missing data. Data were summarized by treatment arm as well as an overall summary.

After removing all observations with missing data, a mixed-effect model was fitted to estimate the overall unadjusted treatment effect for both the change from baseline to short-term MCS and PCS outcomes. This was done so that the subgroups with enhanced treatment effect identified by both the unmodified and modified IPD-SIDES methods can be compared to the overall unadjusted treatment effect to see if the treatment effects actually are more enhanced. The IPD-IT, unmodified IPD-SIDES and modified IPD-SIDES methods were then applied to the pooled dataset for the primary analyses. Two sets of analyses were performed using each of these methods; one for the change from baseline to short-term follow-up MCS and the second for the change from baseline to short-term follow-up PCS. Application of these methods requires

certain parameters to be specified to control or aid the entire tree growing and selection procedure. The minimum number of individuals in any node at any given time was set to 30. The maximum number of levels or splits of the fully grown tree produced by the IPD-IT method was set to 10. The maximum number of levels for the unmodified IPD-SIDES and modified IPD-SIDES method was set to four since there are only four covariates to consider; thus any identified subgroups can only ever be defined by up to four covariates. Moreover, the maximum number of best splits to consider for each node was set the three for the unmodified IPD-SIDES and modified IPD-SIDES methods with a restriction of $p \leq 0.10$ placed on the splitting criterion. Prior to applying both the IPD-SIDES methods, a grid search was performed to select the optimum complexity control parameter sequence for the first four levels of the tree as described in Chapter 6 (section 6.3.2). The grid search was performed considering all permutations of the values from 0 to 1 in steps of 0.2 at the first level, and then 0.2 to 1 in steps of 0.2 at levels 2, 3 and 4. The IPD-IT method used 5-fold cross-validation to evaluate the subtrees and select the final optimum tree. The resampling procedure was repeated 1000 times for both the IPD-SIDES methods to adjust the p-values of the identified candidate subgroups; thus removing any possibly spurious subgroups. The final trees or subgroups identified by the three methods were then summarized.

10.4 Results

Risk of bias assessment of included trials

The risk of bias assessment (Table 10.1) was based on the main results paper of each trial included in the pooled dataset. It was clear that all four trials used the method of randomization to allocate participants. Moreover, all four trials adequately concealed the allocation sequence. None of the trials were double blinded (participant or therapist). All trials adequately reported the completeness of outcome data including

exclusions from the analysis. Three of the four trials showed no signs of selective outcome reporting and provided a description of how the sample size was determined. All but one of the four trials performed the primary analysis on an intention to treat (ITT) basis.

Table 10.1 - Cochrane collaboration Risk of Bias assessment of included trials

	Risk of Bias Assessment (RoBA)*							
	1	2	3	4	5	6	7	8
Trial 1	Y	Y	N	Y	Y	Y	Y	Y
Trial 2	Y	Y	N	Y	Y	Y	Y	Y
Trial 3	Y	Y	N	Y	U	Y	U	N
Trial 4	Y	Y	N	Y	Y	Y	Y	Y

* RoBA, 1. Random sequence generation, 2. Allocation concealment, 3. Blinding, 4. Incomplete outcome data, 5. Selective reporting, 6. Similarity of groups at baseline, 7. Sample size calculation, 8. Intention to treat analysis

Y- Low risk of bias, N- High risk of bias, U- Unclear

Summary of pooled dataset

The pooled IPD dataset consisted of 4,540 individuals in total from four acupuncture trials. Table 10.2 presents a summary of the demographic and baseline data contained within the pooled dataset. The characteristics and baseline values of the individuals seem to be quite similar when comparing the control arm to the intervention arm. Overall, the sample has an average age of 52 with a slightly higher proportion of females (59%) than males (41%). The mean MCS scores and PCS scores are 45.7 (SD: 12.2) and 35.4 (SD: 8.2) respectively.

Summaries of the MCS and PCS component scores and the change from baseline to short-term follow-up MCS and PCS scores are presented in Table 10.3. On average, both

the MCS and PCS scores are greater in the intervention arm at short-term follow-up. Moreover, the mean change from baseline to short-term follow-up MCS and PCS scores are greater in the intervention arm as well. The positive sign for the mean change suggests that those in receipt of the intervention (acupuncture) improve more in terms of both the mental and physical components compared to those in the control arm. This is consistent with published meta-analyses.

Table 10.2 – Summary of demographic and baseline data from pooled dataset

		Control (N=2397)	Intervention (N=2143)	Total (N=4540)
Age	N	2397	2143	4540
	Mean (SD)	52 (13.4)	52 (13.3)	52 (13.3)
	Median	53.0	53.0	53.0
	Range	18 – 88	18-91	18 – 91
	Missing	0	0	0
Gender	Male	981 (41%)	891 (42%)	1872 (41%)
	Female	1416 (59%)	1252 (58%)	2668 (59%)
	Missing	0	0	0
MCS	N	2199	1956	4155
	Mean (SD)	46 (11.9)	45.4 (12.4)	45.7 (12.2)
	Median	48.2	47.2	47.7
	Range	11.3 - 70.5	9.5 - 71.3	9.5 - 71.3
	Missing	198	187	385
PCS	N	2199	1956	4155
	Mean (SD)	35.2 (8.3)	35.7 (8.2)	35.4 (8.2)
	Median	34.5	35.1	34.8
	Range	12.4 - 67.7	13.7 - 61.3	12.4 - 67.7
	Missing	198	187	385

In total, 3845 were included in the final dataset for analyses. The demographic and baseline data were again summarized and observed for the reduced dataset and the covariates were still well balanced across both arms. The overall mean treatment effect estimate for the change from baseline to short-term MCS outcome was 2.50 (SE: 0.342). The overall mean treatment effect estimate for the change from baseline to short-term PCS outcome was 3.75 (SE: 0.282).

IPD-IT results

The IPD-IT method did not identify any moderators of treatment effect when the outcome was the change from baseline to short-term follow-up MCS or when the outcome was the change from baseline to short-term follow-up PCS.

Table 10.3 – Summary of short-term outcome data and change from baseline to short-term follow-up outcome data

		Control (N=2397)	Intervention (N=2143)	Total (N=4540)
MCS	N	2106	1920	4026
	Mean (SD)	47.4 (11.7)	49.1 (11.0)	48.2 (11.4)
	Median	50.4	52.2	51.3
	Range	14.4 - 71.1	8.1 - 69.7	8.1 - 71.1
	Missing	291	223	514
PCS	N	2106	1920	4026
	Mean (SD)	39.0 (9.8)	43.0 (10.0)	40.9 (10.0)
	Median	38.3	43.9	40.6
	Range	13.5 - 69.4	14.0 - 65.9	13.5 - 69.4
	Missing	291	223	514
MCS CHANGE FROM BASELINE	N	2026	1819	3845
	Mean (SD)	1.4 (10.6)	3.7 (10.3)	2.5 (10.5)
	Median	0.5	2.3	1.3
	Range	-38.1 - 44.6	-39.2 - 42.3	-39.2 - 44.6
	Missing	371	324	695
PCS CHANGE FROM BASELINE	N	2026	1819	3845
	Mean (SD)	3.9 (8.2)	7.3 (8.9)	5.5 (8.8)
	Median	2.4	6.1	4
	Range	-23.4 - 33.7	-18.6 - 37.6	-23.4 - 37.6
	Missing	371	324	695

Unmodified IPD-SIDES results

The unmodified IPD-SIDES method found two subgroups when the outcome was the change from baseline to short-term follow-up MCS (Table 10.4). Recall from chapter 7 that the method aims to identify candidate subgroups where each of the candidate subgroups can be thought of as a separate tree. Therefore, the first candidate subgroup or tree suggests that individuals with baseline $MCS \leq 60.5$ ($n=3464$) and the second

candidate subgroup suggests that individuals aged >34 (n=3422) have greater treatment benefit. When the outcome was the change from baseline to short-term follow-up PCS, the unmodified IPD-SIDES method identified 39 subgroups with enhanced treatment effect (Table 10.5). As there are so many candidate subgroups identified in this table, they will not be interpreted here and can be interpreted in the same way as those subgroups identified in Table 10.4. Just to give one example, the third subgroup identified in Table 10.5 suggests that individuals with baseline MCS ≤ 60.5, aged >34 and have baseline PCS > 25.3 have on average a treatment benefit of 3.718 in terms of the physical component score. The fact that so many candidate subgroups were identified makes the overall interpretation of the results quite difficult. Moreover, if required, reducing the number of subgroups to a select few meaningful subgroups could prove to be quite difficult especially when many of the subgroups have similar treatment effects.

Modified IPD-SIDES results

The modified IPD-SIDES method found one subgroup with enhanced treatment effect when the outcome was the change from baseline to short-term follow-up MCS (Table 10.6). Individuals with baseline MCS ≤ 54.5 (n=2701) have greater treatment benefit. When the outcome was the change from baseline to short-term follow-up PCS, the modified IPD-SIDES method identified 4 subgroups with enhanced treatment effect (Table 10.7). Those with a baseline MCS > 51.4 (n=1531) have an average treatment benefit 4.340, those with baseline MCS > 51.4 and baseline PCS ≤ 35.9 (n=919) have an average treatment benefit of 5.436, those with baseline MCS > 54.5 (n=1144) have an average treatment benefit of 4.609 and finally those with Age ≤ 43 at baseline have an average treatment benefit of 4.929.

Unmodified IPD-SIDES results vs. modified IPD-SIDES results

There is a clear distinction between the unmodified IPD-SIDES method and the modified IPD-SIDES method in the subgroups identified when we observe the results. First and foremost, our objective and hence the objective of the modified IPD-SIDES method is to identify subgroups with enhanced treatment effect i.e. subgroups that have a treatment effect that is greater than the overall treatment effect. When we observe the results of the unmodified IPD-SIDES method, we can clearly see that this is not true for many of the subgroups. For example, the first subgroup identified in Table 10.5 has a mean treatment effect of 3.716 which is less than the overall treatment effect (3.75) for the change from baseline to short-term PCS outcome. On the other hand, the results of the modified IPD-SIDES method indicate that all of the identified subgroups actually do have an enhanced treatment effect compared to the overall mean treatment effect of the data. This observation clearly shows the distinction between the two methods and provides positive evidence that the modified IPD-SIDES method does actually identify subgroups with enhanced treatment effect.

Another observation is that many of the treatment effects of the subgroups identified by the unmodified IPD-SIDES method are not that different to their comparator. For example, the first selected subgroup in Table 10.5 has a mean treatment effect of 3.716 and its comparator subgroup has a treatment effect of 3.614. Moreover, there are also subgroups that have been selected that have a treatment effect that is smaller than the comparator subgroup e.g. the second subgroup identified in Table 10.5. As discussed in the previous chapter, the reason this happens is because the original splitting criterion does not directly evaluate the differential effect. For example, the test statistics for the first subgroup in Table 10.5 are $Z_{E1} = \frac{3.716}{0.293} = 12.683$ and $Z_{E2} = \frac{3.614}{0.939} = 3.849$. The difference between the test statistics suggests that there is a big differential effect between the two groups and thus the original splitting criterion will indicate it as being

highly significant. Evaluating the same subgroup using a mixed-effect model with random trial effect, a fixed treatment effect, a fixed binary indicator for the subgroup and a fixed interaction effect, then the test statistic for the interaction effect is 0.240 which is indicative of there being no differential effect present; yet the unmodified IPD-SIDES splitting criterion suggests otherwise. This therefore provides empirical evidence that the new splitting criterion (equation (9.2) in Chapter 9) used by the modified IPD-SIDES method better evaluates the differential effect and does what we require it to do when compared to the original splitting criterion proposed by the authors.

10.5 Discussion

This chapter demonstrated the application of the proposed IPD-IT, unmodified IPD-SIDES and modified IPD-SIDES methods to real individual patient data. There were no subgroups or moderators of treatment effect identified by the IPD-IT method. The difference between the unmodified IPD-SIDES and modified IPD-SIDES methods was clearly demonstrated having applied it to real data. The modified method identified subgroups with enhanced treatment effect and thus demonstrated that it actually does what we require it to do.

The fact that no subgroups were found by the IPD-IT method does not necessarily imply a negative result. Recall from Chapter 3 that one of the reasons for performing exploratory subgroup analyses is to assess the internal consistency of a main effect found in a trial within subgroups. Therefore the fact that no subgroups were found by the IPD-IT method could suggest that the treatment effectiveness found from the primary trial analyses is consistent across subgroups. However, on the other hand,

such a conclusion may not be believable given that the modified IPD-SIDES method did find some subgroups.

The application of these methods to real data brought to light several issues associated with conducting IPD meta-analyses. Although IPD were provided for each of the trials, the data are only of use for subgroup analyses if there are several demographic and baseline covariates that are common to all trials. In the example in this chapter, only four covariates were common to all trials. This may contribute towards the fact that no subgroups were identified by the IPD-IT method because there were not enough covariates that could have been used to define potential subgroups. Moreover, this could also mean that the modified IPD-SIDES method could have missed out on identifying more subgroups defined by a greater number of covariates with larger and more enhanced treatment effect.

Missing data was another issue faced when applying these methods; a problem experienced with most statistical analyses. Observations with missing baseline or outcome data had to be dropped from the analyses. Although there are methods for dealing with missing data e.g. imputation, analysing data with imputed values would only serve as a form of sensitivity analyses.

Typically when performing such analyses, the ICC value is computed beforehand to determine what analyses method to use. The ICC values for the MCS and PCS outcome data were 0.002 and 0.03 respectively. These values are very small, hence the application of the original IT and SIDES methods may have sufficed. To verify this, the original methods were applied to the data and the same results were observed. Nonetheless, the main analyses in this chapter did adjust for clustering effects by

applying the proposed IPD-IT and modified IPD-SIDES methods with the aim of demonstrating the application of these methods to real data.

This chapter concludes the proposal of the IPD-IT and modified IPD-SIDES procedures by demonstrating their application to real individual patient data from several similar trials. The next chapter will thus provide a detailed discussion of the work contained in this thesis along with several recommendations for further work. Finally, the conclusions of this work along with its contributions to LBP research, statistical methodology and medical research will also be presented.

Table 10.4 – Subgroups identified by the unmodified IPD-SIDES method when applied to the change from baseline to short-term MCS outcome data.

Subgroup	Selected subgroup				Disregarded subgroup				Differential effect	Split criterion p-value
	n1	Trt1	SE Trt1	Trt1 p-value	n0	Trt0	SE Trt0	Trt0 p-value		
MCS <= 60.5	3464	2.768	0.361	8.77E-15	381	0.647	0.691	0.1748	2.122	1.93E-06
Age > 34	3422	2.756	0.363	1.60E-14	423	0.013	1.000	0.4947	2.743	8.44E-08

Table 10.5 – Subgroups identified by the unmodified IPD-SIDES method when applied to the change from baseline to short-term PCS outcome data.

Subgroup	Selected subgroup				Disregarded subgroup				Differential effect	Split criterion p-value
	n1	Trt1	SE Trt1	Trt1 p-value	n0	Trt0	SE Trt0	Trt0 p-value		
MCS <= 60.5	3464	3.716	0.293	0.000	381	3.614	0.939	5.95E-05	0.102	4.46E-10
MCS <= 60.5 & Age > 34	3071	3.658	0.307	0.000	393	3.894	0.944	1.87E-05	-0.236	3.72E-08
MCS <= 60.5 & Age > 34 & PCS > 25.3	2755	3.718	0.314	0.000	316	3.597	1.109	0.000666	0.121	1.25E-09
MCS <= 60.5 & Age > 34 & PCS > 28.2	2456	3.552	0.329	0.000	615	4.341	0.738	2.05E-09	-0.789	0.000503
MCS <= 60.5 & Age > 34 & PCS <= 47.1	2764	3.752	0.326	0.000	307	3.166	0.661	8.24E-07	0.585	2.02E-06
MCS <= 60.5 & PCS > 25.5	3114	3.713	0.298	0.000	350	4.163	1.079	5.71E-05	-0.450	1.26E-09
MCS <= 60.5 & PCS > 25.5 & Age <= 63	2519	3.755	0.339	0.000	595	3.470	0.610	6.25E-09	0.285	0.000137
MCS <= 60.5 & PCS > 25.5 & Age > 33	2790	3.740	0.312	0.000	324	3.373	0.988	0.000319	0.367	1.38E-09
MCS <= 60.5 & PCS > 25.5 & Age <= 68	2827	3.721	0.316	0.000	287	3.318	0.861	5.82E-05	0.403	2.07E-08
MCS <= 60.5 & Age <= 69	3154	3.817	0.311	0.000	310	2.369	0.832	0.002197	1.448	2.46E-11
MCS <= 60.5 & Age <= 69 & PCS > 25.8	2839	3.706	0.314	0.000	315	5.249	1.144	2.24E-06	-1.544	3.32E-07
MCS <= 60.5 & Age <= 69 & PCS <= 43.7	2524	4.047	0.355	0.000	630	3.039	0.525	3.53E-09	1.008	7.18E-05

MCS <= 60.5 & Age <= 69 & PCS <= 47.9	2847	3.963	0.328	0.000	307	2.759	0.679	2.45E-05	1.204	1.48E-08
PCS > 25.2	3449	3.909	0.288	0.000	396	3.047	1.000	0.001158	0.862	1.09E-13
PCS > 25.2 & Age <= 68	3106	3.941	0.307	0.000	343	3.361	0.810	1.68E-05	0.580	7.43E-10
PCS > 25.2 & Age <= 68 & MCS > 28.1	2790	3.776	0.323	0.000	316	5.219	0.921	7.39E-09	-1.443	2.14E-05
PCS > 25.2 & Age <= 68 & MCS > 32.8	2484	3.907	0.346	0.000	622	4.058	0.647	1.82E-10	-0.151	0.000368
PCS > 25.2 & Age <= 68 & MCS <= 59.9	2801	3.818	0.320	0.000	305	4.630	0.981	1.19E-06	-0.812	3.52E-07
PCS > 25.2 & Age > 33	3101	3.946	0.301	0.000	348	3.458	0.973	0.000190	0.489	1.49E-11
PCS > 25.2 & Age > 33 & MCS > 28.1	2791	3.826	0.317	0.000	310	4.812	0.909	5.99E-08	-0.986	1.76E-06
PCS > 25.2 & Age > 33 & MCS > 32.9	2469	3.957	0.341	0.000	632	3.903	0.623	1.84E-10	0.055	0.00016
PCS > 25.2 & Age > 33 & MCS <= 60.3	2795	3.812	0.314	0.000	306	4.782	1.002	9.02E-07	-0.970	1.91E-07
PCS > 25.2 & MCS <= 60.2	3112	3.814	0.301	0.000	337	4.321	0.945	2.42E-06	-0.506	1.04E-08
PCS > 25.2 & MCS <= 60.2 & Age > 33	2789	3.825	0.315	0.000	323	3.604	1.009	0.000177	0.222	1.25E-09
PCS > 25.2 & MCS <= 60.2 & Age <= 63	2513	3.899	0.343	0.000	599	3.410	0.604	8.22E-09	0.489	5.32E-05
PCS > 25.2 & MCS <= 60.2 & Age <= 68	2822	3.816	0.319	0.000	290	3.537	0.857	1.85E-05	0.279	3.09E-08
Age <= 69	3480	3.881	0.299	0.000	365	2.329	0.788	0.001550	1.552	1.49E-12
Age <= 69 & PCS > 25.4	3129	3.888	0.304	0.000	351	4.363	1.077	2.56E-05	-0.475	6.11E-10
Age <= 69 & PCS > 25.4 & MCS <= 59.9	2819	3.754	0.317	0.000	310	4.725	0.980	7.14E-07	-0.972	6.94E-07
Age <= 69 & PCS > 25.4 & MCS > 28	2816	3.703	0.320	0.000	313	5.384	0.923	2.76E-09	-1.681	4.93E-05
Age <= 69 & PCS > 25.4 & MCS > 32.8	2503	3.831	0.342	0.000	626	4.083	0.646	1.31E-10	-0.252	0.00055
Age <= 69 & PCS <= 47.6	3137	3.981	0.317	0.000	343	3.056	0.645	1.07E-06	0.925	3.22E-08
Age <= 69 & PCS <= 47.6 & MCS <= 57.3	2522	3.874	0.344	0.000	615	4.234	0.755	1.01E-08	-0.360	6.67E-05
Age <= 69 & PCS <= 47.6 & MCS > 28.8	2816	3.871	0.336	0.000	321	4.793	0.925	1.11E-07	-0.922	7.11E-06
Age <= 69 & PCS <= 47.6 & MCS <= 60.6	2823	3.932	0.330	0.000	314	3.958	1.035	6.53E-05	-0.026	1.10E-08
Age <= 69 & MCS <= 60.4	3135	3.816	0.312	0.000	345	4.061	0.976	1.57E-05	-0.244	1.25E-08
Age <= 69 & MCS <= 60.4 & PCS > 25.9	2813	3.706	0.316	0.000	322	4.986	1.137	5.76E-06	-1.279	2.12E-07
Age <= 69 & MCS <= 60.4 & PCS <= 43.7	2510	4.040	0.357	0.000	625	3.038	0.528	4.41E-09	1.002	8.28E-05
Age <= 69 & MCS <= 60.4 & PCS <= 47.9	2831	3.961	0.330	0.000	304	2.806	0.685	2.09E-05	1.154	2.35E-08

Table 10.6 – Subgroups identified by the modified IPD-SIDES method when applied to the change from baseline to short-term MCS outcome data.

	Selected subgroup				Disregarded subgroup						
Subgroup	n1	Trt1	SE Trt1	Trt1 p-value	n0	Trt0	SE Trt0	Trt0 p-value	SMD* of selected subgroup	Differential effect	Split criterion p-value
MCS <= 54.5	2701	3.285	0.415	1.22E-15	1144	0.635	0.430	0.0700	0.28	2.650	0.0009

* SMD – Standardized mean difference

Table 10.7 – Subgroups identified by the modified IPD-SIDES method when applied to the change from baseline to short-term PCS outcome data.

	Selected subgroup				Disregarded subgroup						
Subgroup	n1	Trt1	SE Trt1	Trt1 p-value	n0	Trt0	SE Trt0	Trt0 p-value	SMD* of selected subgroup	Differential effect	Split criterion p-value
MCS > 51.4	1531	4.340	0.463	0.000	2314	3.403	0.348	0.000	0.56	0.937	0.086
MCS > 51.4 & PCS <= 35.9	919	5.436	0.609	0.000	612	3.054	0.591	0.000	0.70	2.382	0.016
MCS > 54.5	1144	4.609	0.544	0.000	2701	3.378	0.324	0.000	0.59	1.230	0.029
Age <= 43	1170	4.929	0.527	0.000	2675	3.237	0.330	0.000	0.63	1.692	0.005

* SMD – Standardized mean difference

Chapter 11

Discussion, Further Work & Conclusions

11.1 Introduction

In Chapter 8 and Chapter 9, the IPD-IT and modified IPD-SIDES methods were developed and proposed for application in an IPD meta-analysis setting. The previous chapter, Chapter 10, was the final stage of the methodological development demonstrating the application of the proposed IPD-IT and modified IPD-SIDES methods to real life data from several trials of acupuncture for non-specific low back pain. The objective of this chapter is to provide a detailed summary of the work carried out in this thesis. Firstly, a detailed discussion will be presented along with recommendations for further work. Finally, a structured conclusion of the work in this thesis will be presented.

11.2 Discussion and Further Work

The work presented in this thesis contributes towards the advancement of subgroup analyses methodology within the area of LBP. Moreover, it further contributes towards

statistical methodology for identifying subgroups in an IPD meta-analyses framework. The detailed discussion in this section brings to light a number of important research questions that need investigating. These research questions will thus form the recommendations for future work and will also be presented in this section.

Simulation studies

A simulation study was performed in Chapter 7 to evaluate the performance of the IT, STIMA and SIDES methods in a number of very simple scenarios in a single trial setting. The IT and SIDES methods performed well particularly in detecting large and very large interaction effect sizes for a variety of sample sizes. On the other hand, the STIMA method didn't perform well and was therefore dropped from further investigation. In theory the STIMA method sounds quite plausible and one ought to expect that it should perform as described; however this was not the case. Therefore, further work should consider extensively evaluating the STIMA method to investigate why it performed the way it did in the simulation study. Overall, the single trial simulation study provided confidence that the IT and SIDES methods work well in a number of different settings having considered a variety of sample sizes and interaction effect sizes.

A simulation study was also performed in Chapter 8 to evaluate the performance of the two proposed extended methods (IPD-IT and IPD-SIDES) considering a variety of sample sizes, interaction effect sizes and varying between-study variance. The simulation study highlighted an issue with the performance of the IPD-SIDES method when detecting subgroups with very large treatment effects. This issue was thoroughly investigated and resolved in Chapter 9; however, the investigation gave rise to a number of other important findings associated with the method. In particular, the method identifies subgroups with the smallest p-value and thus doesn't actually do what we require it to do. Therefore the method was further developed in Chapter 9 to

do what we require it to do by proposing a new splitting criterion that directly evaluates the interaction effect and also with a restriction imposed on the interaction p-value; hence the modified IPD-SIDES method was proposed. The modified method was then evaluated using a simulation study. Both the IPD-IT and the modified IPD-SIDES methods performed well in detecting interaction effects using either a fixed-effect or mixed-effect model to estimate the splitting criterion to account for the between-study variability. Again, the performance of these methods was assessed in a number of scenarios where the simulated data were generated from a linear model assuming a normally distributed outcome. Having demonstrated the methods work in several simple settings, it would be good to further evaluate their performance under slightly more complex settings such as when the data deviate from normality or when there are imbalances in the data. An advantage of recursive partitioning based methods compared to linear regression is that they do not make any assumptions about the distribution of the outcome variable or any assumptions of linearity. These methods have been shown to perform well in different scenarios when the continuous outcome data are non-normal. For example, Su et al demonstrated in a single trial setting that the IT procedure works well in detecting interactions when there are deviations from normality (136). Therefore, it would be good to evaluate the proposed methods and demonstrate how they perform in the event that data are non-normal. For example, we could evaluate the methods using a simulation study where data are generated as described in Chapter 7 but instead using a non-normal error distribution e.g. an exponential distribution. Other than continuous outcome data, future work should also consider the extension of the methods to other data types such as binary outcome data using a logistic regression model. Moreover, it is also of interest to evaluate the performance of the proposed methods considering varying degrees of imbalance in the data and also considering data with different types of covariates e.g. categorical and continuous. Hence, future work should consider performing a simulation study to

assess the IPD-IT and modified IPD-SIDES methods in more complex scenarios.

Furthermore, it would also be of interest to perform a simulation study to evaluate the performance of the modified IPD-SIDES method using varying restrictions on the splitting criterion p-value. This will help assess how much we can relax the restriction without greatly affecting the performance of the method.

Missing data

The handling of missing data is a limitation of the original IT and SIDES methods as well as the IPD-IT and modified IPD-SIDES methods. Application of these methods to real data requires that the dataset is complete i.e. all observations with missing data need to be removed prior to the analyses. This could result in a loss of a large amount of information if there are several observations with one or more pieces of data missing i.e. missing covariate or outcome data. Not much emphasis was placed on handling missing data in this thesis as the objective of the work was to develop a method(s) and demonstrate its application. One option for handling missing data is to use surrogate splits; an approach proposed by Breiman et al and commonly used by the CART type approach (130, 131). If we assume a best split s^* is found during the tree growing process having ignored all the missing values, then a surrogate split approach basically searches for the next best split s using another covariate that produces a split that is most similar to s^* e.g. it has a similar splitting criterion value. The observation with the missing value is then sent to the node corresponding to its allocation from the surrogate split. Another option to consider is the use of multiple imputation techniques to handle missing data. Whichever approach is taken, analyses of a dataset with imputed values would only serve as a form of sensitivity analyses. However, of the two approaches for handling missing data, it has been shown that multiple imputation outperforms the surrogate split approach in terms of predictive accuracy (168).

Application to empirical data

In Chapter 10, the proposed IPD-IT and modified IPD-SIDES methods were applied to a real IPD dataset consisting of data from four acupuncture trials. The purpose of this was simply to demonstrate the application of the methods. In general for both of these methods, if no subgroups are identified then this does not suggest a negative result. Such a result could be inferred as being confirmation of internal consistency which is one of the reasons for performing exploratory subgroup analyses as discussed in Chapter 3 i.e. confirmation that the treatment effect is consistent across subgroups. However there may be instances, as displayed in the application of the methods in Chapter 10, where one method finds subgroups whereas the other method doesn't. In such an instance, both methods cannot be right. Observation of the simulation study results suggest that the IPD-IT method is more likely to get a false negative than modified IPD-SIDES is to get a false positive, so out of the two, we would probably prefer to believe the modified IPD-SIDES method. Both of these methods were only applied to a single pooled dataset in Chapter 10. However, it would also be nice to demonstrate the application of these methods to other real datasets as well. Hence, further work should consider the application of these methods to more real datasets.

Several trials may test similar interventions using similar outcome measures however they may not have recorded the same or similar baseline measures or demographic data. This would result in very few common covariates being investigated and thus quite likely contributing towards the reason why no subgroups are found. Hence, a lesson learned here is that the commonality of covariates is a very important factor to consider during the planning stages of an IPD meta-analyses in the context of subgroup analyses where the aim is to investigate patient level covariates.

Computational efficiency

Data mining techniques and machine learning algorithms are commonly associated with a reputation of being computationally intensive, in particular when dealing with large datasets. The same is also true for the IPD-IT and modified IPD-SIDES methods. In the simulation studies presented in this thesis, the computational time of these methods was quite short due to only a small number of binary variables being considered. However, when the data contain several more covariates with many more potential split points to consider, the computational time is amplified. This was experienced during the application of the methods to the real dataset in Chapter 10 and is thus noted here as a limitation of these methods. Both of the methods were investigated to see where the time lapse occurs when compared to the original IT and SIDES methods by applying them to the real data in Chapter 10 using the change from baseline to short-term follow-up PCS as the outcome. It was observed that the time taken to evaluate the splitting criterion for a single split was greater for the proposed methods; hence the overall time taken to grow a tree was greater. For example, the original IT method took 6.3 seconds to grow a single tree, the IPD-IT method that uses a fixed-effect model to estimate the splitting criterion took 105.3 seconds to grow a single tree and the IPD-IT method that uses a mixed-effect model to estimate the splitting criterion took 459.1 seconds to grow a single tree. Furthermore, the original SIDES method took 9.3 seconds to grow a single tree, the IPD-SIDES method that uses a fixed-effect model for estimating the splitting criterion took 16.2 seconds and finally the IPD-SIDES method that uses a mixed-effect model took 51.8 seconds to grow a tree. The reason for this is that the splitting criteria of the original methods directly compute the splitting criterion i.e. the t-test statistic, using the data whereas the proposed IPD-IT and modified IPD-SIDES methods fit a model first to account for the clustering and then extract the test statistic thereafter from the fitted model. Direct computation of

the splitting criterion is evidently quicker than obtaining the splitting criterion having fitted a model. As a result, the use of the splitting criterion obtained from a fitted model has a cumulative effect on the time taken for the tree growing process. Consequently, this also lengthens the time taken for any of the cross-validation or resampling components of the methods; thus increasing the overall computational time. This therefore raises a question about whether or not the computational efficiency can be improved for both methods by developing a splitting criterion that can be directly computed without affecting the performance of the methods. One possible option to consider, as briefly discussed in Chapter 8, is to use a two-stage approach by evaluating the original IT and modified SIDES splitting criteria in each trial separately and then synthesizing them using some weighted average as done so in conventional meta-analyses. Although such an approach sounds plausible, it does have an obvious drawback that was also discussed in Chapter 8. Nonetheless, having highlighted the increased computational time of the proposed methods, the two-stage approach is something worth investigating. Therefore, future research should develop and evaluate the performance of the two-stage approach that directly computes the splitting criterion to see if it improves the computational efficiency without any loss of the methods performance.

11.3 General recommendations

Conventional approach vs proposed methods

An IPD meta-analysis framework is ideal for exploring patient level covariates and performing subgroup analyses. The conventional approach to subgroup analyses in this framework involves testing a limited number of subgroups one at a time. Considering the resource intensiveness associated with the entire process of forming an IPD repository, it would be a shame to waste these efforts by just exploring a limited

number of covariates one at a time instead of exploring the entire covariate space to identify subgroups defined by multiple covariates. For this reason, the IPD-IT and IPD-SIDES methods have been developed and proposed in this thesis as two novel tools for subgroup analyses or subgroup identification that make best use of an IPD meta-analysis framework. It is thus recommended that these methods be considered when performing subgroup analyses in an IPD meta-analyses setting.

Choice of method

The choice of which method to use out of the two simply depends on the objective of the researcher. To be more precise, it depends on whether the researcher wants to use the individual patient data to identify moderators of treatment effect or if they want to identify subgroups with enhanced treatment effect i.e. subgroups with a treatment effect greater than the overall treatment effect. Thus, when the research objective is to identify moderators of treatment effect, it is recommended that the IPD-IT method be used. If however, the objective is to identify subgroups with enhanced treatment effect, then it is recommended that the IPD-SIDES method be used.

Splitting criterion estimation

The results of the simulation studies presented in Chapter 8 and Chapter 9 suggest that the performance of the methods are quite similar when using either a fixed-effect or a mixed-effect model to estimate the splitting criterion. Hence, there is no particular preference as to which is used when applying these methods. However, the application of these methods can become computationally intensive when applied to larger datasets with many more variables, as discussed in section 11.2. Therefore, when the datasets are very large with a lot of variables, it is recommended that the methods be applied using a fixed-effect model to estimate the splitting criterion in order to save computational time.

11.4 Conclusions

11.4.1 Addressing the research objectives of this thesis

The primary and secondary research objectives were clearly presented in Chapter 1.

The results and findings presented in this thesis fulfill the primary and secondary research objectives and will now be specified.

Primary research objective

The primary research objective of this thesis was to develop and evaluate innovative approaches for performing subgroup analyses in IPD meta-analyses within the area of low back pain. This thesis satisfies the research objective by proposing two novel approaches to subgroup analyses or subgroup identification, namely the IPD-IT and modified IPD-SIDES methods, in an IPD meta-analyses setting. As mentioned in previous chapters, the aims of these methods are different but are both very important clinically. To be more precise, the IPD-IT method aims to identify subgroups defined by one or more characteristics that are moderators of treatment effect, whereas the objective of the modified IPD-SIDES method is to identify subpopulations defined by one or more characteristics that have enhanced treatment effect. Both of these methods can aid a clinician's decision making allowing them to better target treatments to patients presenting with LBP in the hope to maximize the treatment benefit.

Secondary research objectives

Secondary research objectives were required to systematically work towards answering the primary research objective. This thesis presented a review of the current recommendations for performing subgroup analyses in RCTs in Chapter 3, highlighting some important issues regarding how subgroups are typically defined and

evaluated. A systematic review in Chapter 4, published in the Spine Journal, found that the quality of subgroup analyses in the low back pain literature was of poor quality and that alternative statistical approaches are required. Finally, several alternative statistical methods were described in Chapter 5 from which tree based methods were highlighted as a promising alternative to overcome the associated issues with current guidelines and conventional subgroup analyses. The identified tree methods were thus developed and extended to address the primary research objective.

11.4.2 Contributions to the literature

The work contained in this thesis provides several strengths that contribute to the existing literature. These strengths will now be presented.

Subgroup analyses in the area of LBP

For many years, subgroup analyses have consistently been a key topic of discussion in the LBP research community. Though much effort has been made to undertake subgroup analyses in a single trial setting, the quality of the analyses and the reliability of the results are highly questionable (Chapter 4). Conventional methods used for subgroup analyses in the area of LBP in both a single trial setting and an IPD setting have several issues; one of them being the statistical approach used i.e. testing subgroups individually rather than identifying subgroups defined by multiple characteristics. This thesis identified tree based methods (IT and SIDES) used in other research disciplines as an alternative statistical approach for identifying subgroups defined by multiple characteristics. Hence tree methods provide an alternative approach for performing subgroup analyses in a single LBP trial setting. Moreover, the proposed extensions of the identified methods (IPD-IT and modified IPD-SIDES) now

provide two new novel approaches to subgroup analyses of LBP data in an IPD meta-analyses setting.

Statistical contributions

Subgroup analyses in general in an IPD meta-analyses framework typically test subgroups one at a time using either a one-stage or a two stage-approach. Considering the resource intensiveness of IPD meta-analyses, it would be a shame to collect all of the data only to perform one or a small number of subgroup analyses of pre-specified subgroups with a pre-defined and justified cut-point. Rather, it would be better to make full use of the data and explore the entire covariate space to identify subgroups defined by multiple characteristics. Therefore, this thesis proposes two statistical approaches to subgroup analyses in an IPD meta-analyses setting.

Medical contributions

Although the problem originated from an important research priority in the LBP community, this does not restrict the identified and proposed statistical methods to the field of LBP only. The proposed methods are easily applicable to any field of research where the aim is to perform subgroup analyses in an IPD meta-analyses framework. Identifying subgroups in this manner could be extremely useful in any medical research area so that treatments can be targeted accordingly to maximize benefit and quite possibly to reduce any associated harms.

Bibliography

1. Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine (Phila Pa 1976)*. 2014;39(7):618-29.
2. Andersson GB. Epidemiology of low back pain. *Acta orthopaedica Scandinavica Supplementum*. 1998;281:28-31.
3. Maniadakis N, Gray A. The economic burden of back pain in the UK. *Pain*. 2000;84(1):95-103.
4. Pengel LH, Herbert RD, Maher CG, Refshauge KM. Acute low back pain: systematic review of its prognosis. *BMJ*. 2003;327(7410):323.
5. Schneider S, Schmitt H, Zoller S, Schiltenswolf M. Workplace stress, lifestyle and social factors as correlates of back pain: a representative study of the German working population. *Int Arch Occup Environ Health*. 2005;78(4):253-69.
6. Kent PM, Keating JL. The epidemiology of low back pain in primary care. *Chiropr Osteopat*. 2005;13:13.
7. Steenstra IA, Verbeek JH, Heymans MW, Bongers PM. Prognostic factors for duration of sick leave in patients sick listed with acute low back pain: a systematic review of the literature. *Occup Environ Med*. 2005;62(12):851-60.
8. Thelin A, Holmberg S, Thelin N. Functioning in neck and low back pain from a 12-year perspective: a prospective population-based study. *J Rehabil Med*. 2008;40(7):555-61.
9. Savigny P KS, Watson P, Underwood M, Ritchie G, Cotterell M, Hill D, Browne N, Buchanan E, Coffey P, Dixon P, Drummond C, Flanagan M, Greenough C, Griffiths M, Halliday-Bell J, Hettinga D, Vogel S, Walsh D. Low Back Pain: early management of persistent non-specific low back pain. London: National Collaborating Centre for Primary Care and Royal College of General Practitioners. 2009.

10. Bonica JJ. The need of a taxonomy. *Pain*. 1979;6(3):247-8.
11. Spitzer W, Leblanc FE. Scientific approach to the assessment and management of activity-related spinal disorders. A monograph for clinicians. Report of the Quebec Task Force on Spinal Disorders. *Spine (Phila Pa 1976)*. 1987;12(7 Suppl):S1-59.
12. Underwood MR, Morton V, Farrin A, Team UBT. Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM dataset [ISRCTN32683578]. *Rheumatology (Oxford)*. 2007;46(8):1297-302.
13. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990?2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2012;380(9859):2163-96.
14. Walker BF. The Prevalence of Low Back Pain: A Systematic Review of the Literature from 1966 to 1998. *Journal of Spinal Disorders & Techniques*. 2000;13(3):205-17.
15. Pengel LHM, Herbert RD, Maher CG, Refshauge KM. Acute low back pain: systematic review of its prognosis. *BMJ*. 2003;327(7410):323.
16. Schneider S, Schmitt H, Zoller S, Schiltenswolf M. Workplace stress, lifestyle and social factors as correlates of back pain: a representative study of the German working population. *International Archives of Occupational and Environmental Health*. 2005;78(4):253-69.
17. Macfarlane GJ, Jones GT, Hannaford PC. Managing low back pain presenting to primary care: Where do we go from here? *Pain*. 2006;122(3):219-22.
18. Arthritis Research Campaign. *Arthritis: the big picture*.
19. Waddell G, Feder G, McIntosh A, Lewis M, Hutchinson A. (1996) Low Back Pain Evidence Review London: Royal College of General Practitioners. *Journal of Manual & Manipulative Therapy*. 1998;6(3):151-3.
20. Croft PR, Macfarlane GJ, Papageorgiou AC, Thomas E, Silman AJ. Outcome of low back pain in general practice: a prospective study. *BMJ*. 1998;316(7141):1356.
21. Dagenais S, Caro J, Haldeman S. A systematic review of low back pain cost of illness studies in the United States and internationally. *The Spine Journal*. 2008;8(1):8-20.
22. Airaksinen O, Brox J, Cedraschi C, Hildebrandt J, Klaber-Moffett J, Kovacs F, et al. Chapter 4 European guidelines for the management of chronic nonspecific low back pain. *Eur Spine J*. 2006;15(0):s192-s300.
23. Chou R, Qaseem A, Snow V, Casey D, Cross JT, Shekelle P, et al. Diagnosis and Treatment of Low Back Pain: A Joint Clinical Practice Guideline from the American

College of Physicians and the American Pain Society. *Ann Intern Med*. 2007;147(7):478-91.

24. Deyo RA, Battie M, Beurskens AJ, Bombardier C, Croft P, Koes B, et al. Outcome measures for low back pain research. A proposal for standardized use. *Spine (Phila Pa 1976)*. 1998;23(18):2003-13.

25. Dunn KM, Croft PR. Classification of low back pain in primary care: using "bothersomeness" to identify the most severe cases. *Spine (Phila Pa 1976)*. 2005;30(16):1887-92.

26. Parsons S, Carnes D, Pincus T, Foster N, Breen A, Vogel S, et al. Measuring troublesomeness of chronic pain by location. *BMC Musculoskelet Disord*. 2006;7:34.

27. Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). *Arthritis Care Res (Hoboken)*. 2011;63 Suppl 11:S240-52.

28. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine (Phila Pa 1976)*. 1983;8(2):141-4.

29. Fairbank Jc Fau - Couper J, Couper J Fau - Davies JB, Davies Jb Fau - O'Brien JP, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy*. 1980;66(8):271-3.

30. Von Korff M, Ormel J, Keefe FJ, Dworkin SF. Grading the severity of chronic pain. *Pain*. 1992;50(2):133-49.

31. Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med Care*. 1996;34(3):220-33.

32. Ware JE, Jr. SF-36 health survey update. *Spine (Phila Pa 1976)*. 2000;25(24):3130-9.

33. EuroQol Group. EuroQol-a new facility for the measurement of health-related quality of life. *Health Policy*. 1990;16(3):199-208.

34. Team UBT. United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: effectiveness of physical treatments for back pain in primary care. *BMJ*. 2004;329(7479):1377.

35. Little P, Lewith G, Webley F, Evans M, Beattie A, Middleton K, et al. Randomised controlled trial of Alexander technique lessons, exercise, and massage (ATEAM) for

chronic and recurrent back pain.[Reprint in Br J Sports Med. 2008 Dec;42(12):965-8; PMID: 19096019]. BMJ. 2008;337:a884.

36. Lamb SE, Hansen Z, Lall R, Castelnovo E, Withers EJ, Nichols V, et al. Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis. *Lancet*. 2010;375(9718):916-23.

37. Tilbrook HE, Cox H, Hewitt CE, Kang'ombe AR, Chuang L-H, Jayakody S, et al. Yoga for Chronic Low Back Pain. *Ann Intern Med*. 2011;155(9):569-78.

38. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korf M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine (Phila Pa 1976)*. 2008;33(1):90-4.

39. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

40. Murray CJ, Richards MA, Newton JN, Fenton KA, Anderson HR, Atkinson C, et al. UK health performance: findings of the Global Burden of Disease Study 2010. *Lancet*. 2013;381(9871):997-1020.

41. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*. 2005;365(9454):176-86.

42. Wang R, Ware JH. Detecting Moderator Effects Using Subgroup Analyses. *Prevention science : the official journal of the Society for Prevention Research*. 2011.

43. Lagakos SW. The Challenge of Subgroup Analyses — Reporting without Distorting. *New England Journal of Medicine*. 2006;354(16):1667-9.

44. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. *New England Journal of Medicine*. 2007;357(21):2189-94.

45. Henschke N, Ostelo RW, van Tulder MW, Vlaeyen JW, Morley S, Assendelft WJ, et al. Behavioural treatment for chronic low-back pain. *The Cochrane database of systematic reviews*. 2010(7):CD002014.

46. Hayden JA, van Tulder MW, Malmivaara A, Koes BW. Exercise therapy for treatment of non-specific low back pain. *The Cochrane database of systematic reviews*. 2005(3):CD000335.

47. Walker BF, French SD, Grant W, Green S. Combined chiropractic interventions for low-back pain. *The Cochrane database of systematic reviews*. 2010(4):CD005427.

48. Engers A, Jellema P, Wensing M, van der Windt DA, Grol R, van Tulder MW. Individual patient education for low back pain. *The Cochrane database of systematic reviews*. 2008(1):CD004057.

49. Furlan AD, Brosseau L, Imamura M, Irvin E. Massage for low-back pain: a systematic review within the framework of the Cochrane Collaboration Back Review Group. *Spine (Phila Pa 1976)*. 2002;27(17):1896-910.
50. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials. *JAMA: The Journal of the American Medical Association*. 1991;266(1):93-8.
51. Sibbald B, Roland M. Understanding controlled trials: Why are randomised controlled trials important? *BMJ*. 1998;316(7126):201.
52. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Statistics in Medicine*. 1984;3(4):409-20.
53. Collins R PR, Gray R, Parish S. Large-scale randomised evidence: trials and overviews. In: DJ Weatherall, JGG Ledingham and DA Warrell, Editors, *Oxford Textbook of Medicine* Oxford, Oxford University Press, London. 1996:21-32.
54. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*. 2002;21(19):2917-30.
55. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet*. 2000;355(9209):1064-9.
56. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. *Archives of general psychiatry*. 2002;59(10):877-83.
57. Matthews JNS, Altman DG. Statistics Notes: Interaction 2: compare effect sizes not P values. *BMJ*. 1996;313(7060):808.
58. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. *Journal of Clinical Epidemiology*. 2004;57(3):229-36.
59. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health technology assessment (Winchester, England)*. 2001;5(33):1-56.
60. Altman DG, Matthews JNS. Statistics Notes: Interaction 1: heterogeneity of effects. *BMJ*. 1996;313(7055):486.
61. Randomised trial of intravenous streptokinase, oral aspirin, both or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. *The Lancet*. 1988;332(8607):349-60.

62. Bypass Angioplasty Revascularisation Investigation (BARI). Comparison of Coronary Bypass Surgery with Angioplasty in Patients with Multivessel Disease. *New England Journal of Medicine*. 1996;335(4):217-25.
63. Proschan MA, Waclawiw MA. Practical Guidelines for Multiplicity Adjustment in Clinical Trials. *Controlled Clinical Trials*. 2000;21(6):527-39.
64. Bender R, Lange S. Adjusting for multiple testing--when and how? *Journal of Clinical Epidemiology*. 2001;54(4):343-9.
65. Marshall SW. Power for tests of interaction: effect of raising the Type I error rate. *Epidemiologic perspectives & innovations* : EP+I. 2007;4:4.
66. Casey PD. Does Raising Type 1 Error Rate Improve Power to Detect Interactions in Linear Regression Models? A Simulation Study. *PLoS ONE*. 2013;8(8).
67. Klebanoff MA. Subgroup analysis in obstetrics clinical trials. *Am J Obstet Gynecol*. 2007;197(2):119-22.
68. Kent D, Rothwell P, Ioannidis J, Altman D, Hayward R. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*. 2010;11(1):85.
69. Pincus T, Miles C, Froud R, Underwood M, Carnes D, Taylor S. Methodological criteria for the assessment of moderators in systematic reviews of randomised controlled trials: a consensus study. *BMC Medical Research Methodology*. 2011;11(1):14.
70. Kent P, Keating J, Leboeuf-Yde C. Research methods for subgrouping low back pain. *BMC Medical Research Methodology*. 2010;10(1):62.
71. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340.
72. Pincus T, Miles C, Froud R, Underwood M, Carnes D, Taylor SJ. Methodological criteria for the assessment of moderators in systematic reviews of Randomised Controlled Trials: a consensus study. *BMC Med Res Methodol*. 2011;11(1):14-.
73. Sheets C, Machado LA, Hancock M, Maher C. Can we predict response to the McKenzie method in patients with acute low back pain? A secondary analysis of a randomized controlled trial. *Eur Spine J*. 2012;21(7):1250-6.
74. Smeets RJEM, Maher CG, Nicholas MK, Refshauge KM, Herbert RD. Do psychological characteristics predict response to exercise and advice for subacute low back pain? *Arthritis Rheum*. 2009;61(9):1202-9.
75. Underwood M, Mistry D, Lall R, Lamb S. Predicting response to a cognitive-behavioral approach to treating low back pain: Secondary analysis of the BeST data set. *Arthritis Care & Research*. 2011;63(9):1271-9.

76. Witt CM, Jena S, Selim D, Brinkhaus B, Reinhold T, Wruck K, et al. Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. *Am J Epidemiol.* 2006;164(5):487-96.
77. Gudavalli MR, Cambron JA, McGregor M, Jedlicka J, Keenum M, Ghanayem AJ, et al. A randomized clinical trial and subgroup analysis to compare flexion-distraction with active exercise for chronic low back pain. *Eur Spine J.* 2006;15(7):1070-82.
78. Hansen FR, Bendix T, Skov P, Jensen CV, Kristensen JH, Krohn L, et al. Intensive, dynamic back-muscle exercises, conventional physiotherapy, or placebo-control treatment of low-back pain. A randomized, observer-blind trial. *Spine.* 1993;18(1):98-108.
79. Hay EM, Mullis R, Lewis M, Vohora K, Main CJ, Watson P, et al. Comparison of physical treatments versus a brief pain-management programme for back pain in primary care: a randomised clinical trial in physiotherapy practice. *Lancet.* 2005;365(9476):2024-30.
80. Johnson RE, Jones GT, Wiles NJ, Chaddock C, Potter RG, Roberts C, et al. Active exercise, education, and cognitive behavioral therapy for persistent disabling low back pain: a randomized controlled trial. *Spine.* 2007;32(15):1578-85.
81. Bishop MD, Bialosky JE, Cleland JA. Patient expectations of benefit from common interventions for low back pain and effects on outcome: secondary analysis of a clinical trial of manual therapy interventions. *The Journal of manual & manipulative therapy.* 2011;19(1):20-5.
82. Beurskens AJ, de Vet HC, Koke AJ, Lindeman E, Regtop W, van der Heijden GJ, et al. Efficacy of traction for non-specific low back pain: a randomised clinical trial. *Lancet.* 1995;346(8990):1596-600.
83. Cherkin DC, Deyo RA, Battie M, Street J, Barlow W. A comparison of physical therapy, chiropractic manipulation, and provision of an educational booklet for the treatment of patients with low back pain. *New England Journal of Medicine.* 1998;339(15):1021-9.
84. Hsieh LL-C, Kuo C-H, Yen M-F, Chen TH-H. A randomized controlled clinical trial for low back pain treated by acupressure and physical therapy. *Prev Med.* 2004;39(1):168-76.
85. Roche G, Ponthieux A, Parot-Shinkel E, Jousset N, Bontoux L, Dubus V, et al. Comparison of a functional restoration program with active individual physical therapy for patients with chronic low back pain: a randomized controlled trial. *Archives of Physical Medicine & Rehabilitation.* 2007;88(10):1229-35.

86. Seferlis T, Nemeth G, Carlsson AM, Gillstrom P. Conservative treatment in patients sick-listed for acute low-back pain: a prospective randomised study with 12 months' follow-up. *Eur Spine J*. 1998;7(6):461-70.
87. Bendix AF, Bendix T, Hastrup C. Can it be predicted which patients with chronic low back pain should be offered tertiary rehabilitation in a functional restoration program? A search for demographic, socioeconomic, and physical predictors. *Spine*. 1998;23(16):1775-83; discussion 83-4.
88. Mellin G, Hurri H, Harkapaa K, Jarvikoski A. A controlled study on the outcome of inpatient and outpatient treatment of low back pain. Part II. Effects on physical measurements three months after treatment. *Scand J Rehabil Med*. 1989;21(2):91-5.
89. Carr JL, Klaber Moffett JA, Howarth E, Richmond SJ, Torgerson DJ, Jackson DA, et al. A randomized trial comparing a group exercise programme for back pain patients with individual physiotherapy in a severely deprived area. *Disabil Rehabil*. 2005;27(16):929-37.
90. Jellema P, van der Windt DAWM, van der Horst HE, Blankenstein AH, Bouter LM, Stalman WAB. Why is a treatment aimed at psychosocial factors not effective in patients with (sub)acute low back pain? *Pain*. 2005;118(3):350-9.
91. Klaber Moffett JAMPM, Carr JM, Howarth EM. High Fear-Avoiders of Physical Activity Benefit From an Exercise Program for Patients With Back Pain. *Spine*. 2004;29(11):1167-72.
92. van der Roer N, van Tulder M, Barendse J, Knol D, van Mechelen W, de Vet H. Intensive group training protocol versus guideline physiotherapy for patients with chronic low back pain: a randomised controlled trial. *Eur Spine J*. 2008;17(9):1193-200.
93. Vollenbroek-Hutten MMR, Hermens HJ, Wever D, Gorter M, Rinket J, Ijzerman MJ. Differences in outcome of a multidisciplinary treatment between subgroups of chronic low back pain patients defined using two multiaxial assessment instruments: the multidimensional pain inventory and lumbar dynamometry. *Clin Rehabil*. 2004;18(5):566-79.
94. Becker A, Leonhardt C, Kochen MM, Keller S, Wegscheider K, Baum E, et al. Effects of two guideline implementation strategies on patient outcomes in primary care: a cluster randomized controlled trial. *Spine*. 2008;33(5):473-80.
95. Juni P, Battaglia M, Nuesch E, Hammerle G, Eser P, van Beers R, et al. A randomised controlled trial of spinal manipulative therapy in acute low back pain. *Ann Rheum Dis*. 2009;68(9):1420-7.

96. Smeets RJE, Vlaeyen JWS, Hidding A, Kester ADM, van der Heijden GJMG, van Geel ACM, et al. Active rehabilitation for chronic low back pain: cognitive-behavioral, physical, or both? First direct post-treatment results from a randomized controlled trial [ISRCTN22714229]. *BMC Musculoskelet Disord.* 2006;7:5.
97. Thomas KJ, MacPherson H, Thorpe L, Brazier J, Fitter M, Campbell MJ, et al. Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. *BMJ.* 2006;333(7569):623.
98. Ferreira ML, Ferreira PH, Latimer J, Herbert RD, Maher C, Refshauge K. Relationship between spinal stiffness and outcome in patients with chronic low back pain. *Manual Therapy.* 2009;14(1):61-7.
99. Sherman KJ, Cherkin DC, Ichikawa L, Avins AL, Barlow WE, Khalsa PS, et al. Characteristics of patients with chronic back pain who benefit from acupuncture. *BMC Musculoskelet Disord.* 2009;10:114.
100. van der Hulst M, Vollenbroek-Hutten MMR, Groothuis-Oudshoorn KGM, Hermens HJ. Multidisciplinary rehabilitation treatment of patients with chronic low back pain: a prognostic model for its outcome. *Clin J Pain.* 2008;24(5):421-30.
101. Cecchi F, Negrini S, Pasquini G, Paperini A, Conti AA, Chiti M, et al. Predictors of functional outcome in patients with chronic low back pain undergoing back school, individual physiotherapy or spinal manipulation. *Eur J Phys Rehabil Med.* 2012;48(3):371-8.
102. Cherkin DC, Eisenberg D, Sherman KJ, Barlow W, Kaptchuk TJ, Street J, et al. Randomized trial comparing traditional Chinese medical acupuncture, therapeutic massage, and self-care education for chronic low back pain. *Arch Intern Med.* 2001;161(8):1081-8.
103. Cherkin DC, Sherman KJ, Avins AL, Erro JH, Ichikawa L, Barlow WE, et al. A randomized trial comparing acupuncture, simulated acupuncture, and usual care for chronic low back pain. *Arch Intern Med.* 2009;169(9):858-66.
104. Kalauokalani D, Cherkin DC, Sherman KJ, Koepsell TD, Deyo RA. Lessons from a trial of acupuncture and massage for low back pain: patient expectations and treatment effects. *Spine.* 2001;26(13):1418-24.
105. Karjalainen K, Malmivaara A, Mutanen P, Roine R, Hurri H, Pohjolainen T. Mini-intervention for subacute low back pain: two-year follow-up and modifiers of effectiveness. *Spine.* 2004;29(10):1069-76.
106. Kole-Snijders AM, Vlaeyen JW, Goossens ME, Rutten-van Molken MP, Heuts PH, van Breukelen G, et al. Chronic low-back pain: what does cognitive coping skills

- training add to operant behavioral treatment? Results of a randomized clinical trial. *J Consult Clin Psychol*. 1999;67(6):931-44.
107. Myers SS, Phillips RS, Davis RB, Cherkin DC, Legedza A, Kaptchuk TJ, et al. Patient expectations as predictors of outcome in patients with acute low back pain. *Journal of General Internal Medicine*. 2008;23(2):148-53.
 108. Smeets RJEM, Vlaeyen JWS, Hidding A, Kester ADM, van der Heijden GJMG, Knottnerus JA. Chronic low back pain: physical training, graded activity with problem solving training, or both? The one-year post-treatment results of a randomized controlled trial.[Reprint in *Ned Tijdschr Geneesk*. 2009 Mar 21;153(12):543-9; PMID: 19368107]. *Pain*. 2008;134(3):263-76.
 109. Lachenbruch PA. A note on sample size computation for testing interactions. *Statistics in Medicine*. 1988;7(4):467-9.
 110. Pincus T, Santos R, Breen A, Burton AK, Underwood M. A review and proposal for a core set of factors for prospective cohorts in low back pain: A consensus statement. *Arthritis Care & Research*. 2008;59(1):14-24.
 111. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ*. 2009;338.
 112. Hastie T, Tibshirani R. *Generalized Additive Models*: Chapman and Hall; 1990.
 113. Peck LR. Using Cluster Analysis in Program Evaluation. *Evaluation Review*. 2005;29(2):178-96.
 114. Klonsky ED, Olino TM. Identifying clinically distinct subgroups of self-injurers among young adults: a latent class analysis. *J Consult Clin Psychol*. 2008;76(1):22-7.
 115. Neumann M, Wirtz M, Ernstmann N, Ommen O, Langler A, Edelhauser F, et al. Identifying and predicting subgroups of information needs among cancer patients: an initial study using latent class analysis. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*. 2011;19(8):1197-209.
 116. Schemper M. Non-parametric analysis of treatment—covariate interaction in the presence of censoring. *Statistics in Medicine*. 1988;7(12):1257-66.
 117. Shuster J, van Eys J. Interaction between prognostic factors and treatment. *Controlled Clinical Trials*. 1983;4(3):209-14.
 118. Byar DP. Assessing apparent treatment—covariate interactions in randomized clinical trials. *Statistics in Medicine*. 1985;4(3):255-63.
 119. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*. 1985;41(2):361-72.
 120. Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. *Stat Med*. 1993;12(13):1239-48.

121. Li J, Chan IS. Detecting qualitative interactions in clinical trials: an extension of range test. *Journal of biopharmaceutical statistics*. 2006;16(6):831-41.
122. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Nonparametric Tests for Treatment Effect Heterogeneity. SSRN eLibrary. 2006.
123. Kraemer HC. Toward non-parametric and clinically meaningful moderators and mediators. *Statistics in Medicine*. 2008;27(10):1679-92.
124. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080.
125. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006;25(1):127-41.
126. Royston P, Sauerbrei W. Two techniques for investigating interactions between treatment and continuous covariates in clinical trials. *Stata Journal*. 2009;9(2):230-51.
127. Bonetti M, Gelber RD. A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. *Statistics in Medicine*. 2000;19(19):2595-609.
128. Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. *Clinical trials (London, England)*. 2011;8(2):129-43.
129. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol*. 2006;6:18.
130. Heping Zhang, Singer BH. *Recursive Partitioning and Applications*, 2nd Edition. 2 ed: Springer; 2010.
131. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees.*: Chapman & Hall; 1984.
132. Dusseldorp E, Meulman JJ. The Regression Trunk Approach to Discover Treatment Covariate Interaction. *Psychometrika*. 2004;69:p355-74.
133. Dusseldorp E, Conversano C, Van Os BJ. Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics*. 2010(Advanced online publication).
134. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*. 2011;30(24):2867-80.
135. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*. 2011;30(21):2601-21.

136. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup Analysis via Recursive Partitioning. SSRN eLibrary. 2009.
137. Sutton AJ, Higgins JP. Recent developments in meta-analysis. *Stat Med*. 2008;27(5):625-50.
138. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. *Journal of American Statistical Association*. 1963;58:415-34.
139. Friedman JH. Multivariate Adaptive Regression Splines. 1991:1-67.
140. Quinlan JR, editor *Learning with Continuous {C}lasses*. 5th Australian Joint Conference on Artificial Intelligence; 1992.
141. Loh W-Y. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2011;1(1):14-23.
142. Breiman L. Technical Note: Some Properties of Splitting Criteria. *Machine Learning*. 1996;24(1):41-7.
143. Leblanc M, Crowley J. Survival Trees by Goodness of Split. *Journal of the American Statistical Association*. 1993;88(422):457-67.
144. Doyle P. The Use of Automatic Interaction Detector and Similar Search Procedures. *Operational Research Quarterly (1970-1977)*. 1973;24(3):465-7.
145. Shih Y-S, Tsai H-W. Variable selection bias in regression trees with constant fits. *Computational Statistics & Data Analysis*. 2004;45(3):595-607.
146. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical trials (London, England)*. 2005;2(3):209-17.
147. Abdoell M, LeBlanc M, Stephens D, Harrison RV. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Stat Med*. 2002;21(22):3395-409.
148. Keon Lee S. On generalized multivariate decision tree by using GEE. *Computational Statistics & Data Analysis*. 2005;49(4):1105-19.
149. Lee S. On Classification and Regression Trees for Multiple Responses. In: Banks D, McMorris F, Arabie P, et al., editors. *Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation*: Springer Berlin Heidelberg; 2004. p. 177-84.
150. Loh W-Y, Zheng W. Regression trees for longitudinal and multiresponse data. 2013:495-522.
151. Sela R, Simonoff J. RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*. 2012;86(2):169-207.

152. Su X, Meneses K, McNees P, Johnson WO. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2011;60(3):457-74.
153. Hajjem A, Bellavance F, Larocque D. Mixed effects regression trees for clustered data. *Statistics & Probability Letters*. 2011;81(4):451-9.
154. Greenland S. Principles of multilevel modelling. *International Journal of Epidemiology*. 2000;29(1):158-67.
155. Peugh JL. A practical guide to multilevel modeling. *Journal of School Psychology*. 2010;48(1):85-112.
156. Abo-Zaid G, Guo B, Deeks JJ, Debray TP, Steyerberg EW, Moons KG, et al. Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol*. 2013;66(8):865-73.e4.
157. Higgins JP, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. *Stat Med*. 2001;20(15):2219-41.
158. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat Med*. 2008;27(11):1870-93.
159. Whitehead A, Omar RZ, Higgins JP, Savaluny E, Turner RM, Thompson SG. Meta-analysis of ordinal outcomes using individual patient data. *Stat Med*. 2001;20(15):2243-60.
160. Pinheiro J, Bates D. *Mixed-effects models in S and S-PLUS*: New York: Springer; 2000.
161. Brown HK, Kempton RA. The application of REML in clinical trials. *Stat Med*. 1994;13(16):1601-17.
162. Fisher DJ, Copas AJ, Tierney JF, Parmar MK. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *J Clin Epidemiol*. 2011;64(9):949-67.
163. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics*. 1946;2(6):110-4.
164. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 1997;53(3):983-97.
165. Hempel S, Miles JN, Booth MJ, Wang Z, Morton SC, Shekelle PG. Risk of bias: a simulation study of power to detect study-level moderator effects in meta-analysis. *Systematic reviews*. 2013;2:107.

166. Vickers A. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Medical Research Methodology*. 2001;1(1):1-4.
167. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj*. 2011;343:d5928.
168. Feelders A. Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation? In: Żytkow J, Rauch J, editors. *Principles of Data Mining and Knowledge Discovery. Lecture Notes in Computer Science*. 1704: Springer Berlin Heidelberg; 1999. p. 329-34.

PART IV

Appendices

Appendix A

Appendix A: Systematic review paper published in the Spine Journal

LITERATURE REVIEW

Evaluating the Quality of Subgroup Analyses in Randomized Controlled Trials of Therapist-Delivered Interventions for Nonspecific Low Back Pain

A Systematic Review

Dipesh Mistry, MSc, Shilpa Patel, MSc, BSc Hons, C.Psychol, Siew Wan Hee, PhD, Nigel Stallard, PhD, and Martin Underwood, MD

Study Design. Systematic review.

Objective. To evaluate the quality, conduct, and reporting of subgroup analyses performed in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain (NSLBP).

Summary of Background Data. Randomized controlled trials of therapist-delivered interventions for NSLBP to date have, at best, shown small to moderate positive effects. Identifying subgroups is an important research priority. This review evaluates the quality, conduct, and reporting of subgroup analyses performed in the NSLBP literature.

Methods. Multiple electronic databases were searched for randomized controlled trials of therapist-delivered interventions for NSLBP. Of the identified articles, only articles reporting subgroup analyses (confirmatory or exploratory) were included in the final review. Methodological criteria were used to evaluate the quality of subgroup analyses. The quality of conduct and reporting was also evaluated.

Results. Thirty-nine articles were included in the final review. Of these, only 3 (8%) tested hypotheses about moderators (confirmatory findings), 18 (46%) generated hypotheses about moderators to inform future research (exploratory findings), and 18 (46%) provided insufficient findings. The appropriate statistical test for interaction

was performed in 27 of the articles, of which 10 reported results from interaction tests, 4 incorrectly reported results within individual subgroups, and the remaining articles reported either *P* values or nothing at all.

Conclusion. Subgroup analyses performed in NSLBP trials have been severely underpowered, are only able to provide exploratory or insufficient findings, and have rather poor quality of reporting. Using current approaches, few definitive trials of subgrouping in back pain are very likely to be performed. There is a need to develop new approaches to subgroup identification in back pain research.

Key words: low back pain, back pain, randomized controlled trial, subgroup analysis, effect modification, heterogeneity of treatment effect, interaction.

Level of Evidence: 1

Spine 2014;39:618-629

Low back pain is a common and costly health condition.^{1,2} In the United Kingdom, it affects nearly a third of all adults. Globally, it contributes the most to the overall years lived with disability.³ During recent years, major investments have been made in conducting randomized controlled trials (RCTs) to evaluate various interventions for the treatment of nonspecific low back pain (NSLBP), in particular therapist-delivered interventions. Interventions with some evidence of effectiveness include exercise, yoga, acupuncture, manipulation, postural approaches, and psychological approaches.⁴⁻⁸ At best, the mean effect sizes found in definitive trials are small to moderate. Effects of such magnitude may not be important at an individual patient level but can have important public health benefits. The identification of subgroups that gain the most benefit from interventions for the management of NSLBP is, however, an important research priority internationally.^{4,9-11} Advancing this area of research could have important implications for current clinical practice, thus improving individual patient care.

Typically, RCTs are designed and powered to test a main hypothesis in terms of a primary outcome measure. Any secondary subgroup analyses are likely to be substantially

From the Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, United Kingdom.

Acknowledgment date: July 12, 2013. Revision date: August 2, 2013. Acceptance date: January 13, 2014.

The manuscript submitted does not contain information about medical device(s)/drug(s).

The National Institute for Health Research, under its Programme Grants for Applied Research (RP-PG-0608-10076), funds were received in support of this work. This project benefited from facilities funded through Birmingham Science City Translational Medicine Clinical Research and Infrastructure Trials platform, with support from Advantage West Midlands.

Relevant financial activities outside the submitted work: board membership, expert testimony, grants, payment for lecture, stocks.

Address correspondence and reprint requests to Dipesh Mistry, MSc, Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Gibbet Hill Rd, Coventry, CV4 7AL; E-mail: D.Mistry@Warwick.ac.uk

DOI: 10.1097/BRS.0000000000000231

618 www.spinejournal.com

Copyright © 2014 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

April 2014

underpowered to detect the same effect size as that of the primary analysis. Notwithstanding this lack of statistical power, secondary subgroup analyses of existing trial data have the potential to identify important patient subgroups, provided that the quality of subgroup analyses is of a high standard. Any identified subgroups would need to be further investigated in future trials. Numerous articles in the low back pain literature claim to have performed subgroup analyses; however, the overall quality of these analyses is unknown. We have therefore reviewed the quality, conduct, and reporting of the analyses performed to allow interpretation of their findings.

This article reports a systematic review of RCTs of therapist-delivered interventions for NSLBP that report subgroup analyses (both confirmatory and exploratory) and assess the quality of these analyses.

MATERIALS AND METHODS

Search Strategy

We searched MEDLINE (1948 to July 2013), Ovid MEDLINE(R) In-process & Other Non-Indexed Citations, EMBASE (1974 to July 2013), Web of Science and Citation Index, and Cochrane Controlled Trials Register (CENTRAL). We sought to identify all articles reporting RCTs of therapist-delivered interventions for NSLBP that performed subgroup analyses. For that reason, we initially searched for “low back pain” terms, “RCT” terms, and key subgroup analyses terms such as “subgroup,” “effect modifier,” and “moderator.” These search terms identified only articles that had subgroup analyses terms in the title or abstract and therefore missed out publications that had these terms only in the main text. We therefore reran our searches to identify all RCTs of therapist-delivered interventions in the area of NSLBP using only the key words for “low back pain” and “RCTs.” A full list of the search terms used can be found in the Appendix (available at <http://links.lww.com/BRS/A853>).

Selection of Articles

We scanned all of the titles and abstracts from the search results retrieved to identify all articles potentially reporting a subgroup analysis of an RCT testing a therapist-delivered intervention for NSLBP. We then examined the full text of articles to see whether they performed some form of subgroup analyses and then using specific inclusion and exclusion criteria we decided which articles to include in the final review (Table 1).

Quality of Subgroup Analysis

Articles included in the final review were assessed for the quality of subgroup analyses using the Pincus criteria¹²:

1. Was the subgroup analysis specified *a priori*?
2. Was the selection of subgroup factors for analysis theory/evidence driven?
3. Were subgroup factors measured prior to randomization?
4. Was measurement of subgroup factors measured by adequate (reliable and valid) measurements, appropriate for the target population?

TABLE 1. Inclusion and Exclusion Criteria Used to Select Articles

Inclusion criteria
Randomized controlled trials
Participants 18 yr or older with a history of NSLBP
Therapist-delivered interventions for NSLBP (including psychological interventions and intensive rehabilitation programs)
Primary or secondary analysis of RCTs reporting that a subgroup analysis had been conducted
Exclusion criteria
LBP with known likely cause (fracture, infection, malignancy-specific cause, ankylosing spondylitis, and other inflammatory disorders)
Studies investigating disorders in addition to NSLBP, e.g., NSLBP and neck pain
Outcome not a valid clinical measure of NSLBP, e.g., number of days for sick leave
Testing a clinical prediction rule
Treatment effect modification over time, i.e., Treatment × Moderator × Time
Pooled data sets of similar trials
<i>NSLBP indicates nonspecific low back pain; RCT, randomized controlled trial; LBP, low back pain.</i>

5. Does the analysis contain an explicit test of the interaction between moderator and treatment?

The quality assessment classifies articles as either providing confirmatory findings or exploratory findings. Confirmatory findings support hypotheses about moderators (hypothesis testing), whereas exploratory findings inform future research (hypothesis generating). If an article satisfies all 5 criteria, then its findings are regarded as being confirmatory. Articles satisfying only criteria 3, 4, and 5 are regarded as having exploratory findings. All other remaining articles are regarded as having insufficient findings. The quality of subgroup analyses in all of the identified articles was assessed separately by 3 independent reviewers (D.M., S.P., and S.W.H.). Any discrepancies at the end of the process were resolved through discussion.

Assessment of Conduct and Reporting

There are a number of proposed guidelines that exist for the conduct and reporting of subgroup analyses to ensure that the conclusions drawn are plausible.^{13,14} These guidelines were used to help evaluate the conduct and reporting in the articles identified from the literature search. In particular, 3 areas were assessed: (1) design and methods; (2) reporting of results; and (3) interpretation and discussion.

The design and methods were assessed for all articles, whereas the reporting of results and the interpretation and discussion were assessed only for those articles that used interaction tests

TABLE 2. Key Recommendations in the Area of Subgroup Analyses

Exact subgroup definitions should be given beforehand for continuous and categorical variables along with some justification to avoid <i>post hoc</i> data-dependent definitions of subgroups.
Subgroup analyses should be performed on the primary outcome in the study. This is simply because trials are designed to detect differences only in the primary outcome; therefore, performing subgroup analyses on any other outcome measure will substantially reduce the power.
A differential subgroup effect should be formally evaluated using a statistical test for interaction and the interaction effect reported. Performing tests within individual subgroups and then comparing the results provide an incorrect approach to performing subgroup analyses because it does not directly evaluate the subgroup effect.
The number of subgroup analyses to be performed should be kept to a minimum. This is to avoid the issue of false-positive discovery (type I error inflation) due to multiple testing, a well-known issue if there are several subgroups of interest. Any concerns regarding multiplicity should be acknowledged and addressed appropriately, e.g., applying a Bonferroni correction.

for subgroup analyses. The conduct and reporting of all articles were examined to see whether they conformed to 4 key recommendations in the area of subgroup analyses (Table 2).^{13,14}

RESULTS

Selection of Articles

We identified 4933 articles from the screening of titles and abstracts, of which 4873 articles were excluded. The full texts

for the remaining 60 articles were examined, and an additional 21 articles were excluded on the basis of the inclusion and exclusion criteria. The remaining 39 articles were included in the final review (Figure 1, Tables 3 and 4).

Most of the included articles were of studies conducted in the Netherlands (n = 8; 21%), UK (n = 8; 21%), or the United States (n = 8; 21%). The median study size of the included articles was 223 (range, 100–3093). Twenty-nine articles (74%) performed subgroup analyses on a total study size of around 300 or fewer; the remaining 10 articles (26%) had more than 400 patients.

Quality of Subgroup Analyses

Of the 39 articles, 3 articles (8%) met all 5 criteria and therefore provided confirmatory findings^{15–17}; 18 articles (46%) provided exploratory findings, that is, they met criteria 3, 4, and 5; and 18 articles (46%) provided insufficient findings (Table 3).

Assessment of Conduct and Reporting: Design and Methods

Only one trial was designed to have adequate power to detect important treatment-covariate interactions; however, no specific subgroups of interest were specified *a priori*.¹⁸ The majority of the articles (n = 31; 79%) did not prespecify which subgroups were to be investigated in the analyses. Eight articles prespecified subgroups for confirmatory analyses.^{15–17,19–23} Six of these additionally performed exploratory analyses; however, this distinction was not always made clear at the outset. Baseline characteristics that were prespecified for subgroup analyses included age, sex, baseline Roland and Morris Disability Questionnaire score, psychological distress, workload, history of back pain, radiculopathy, patient preference,

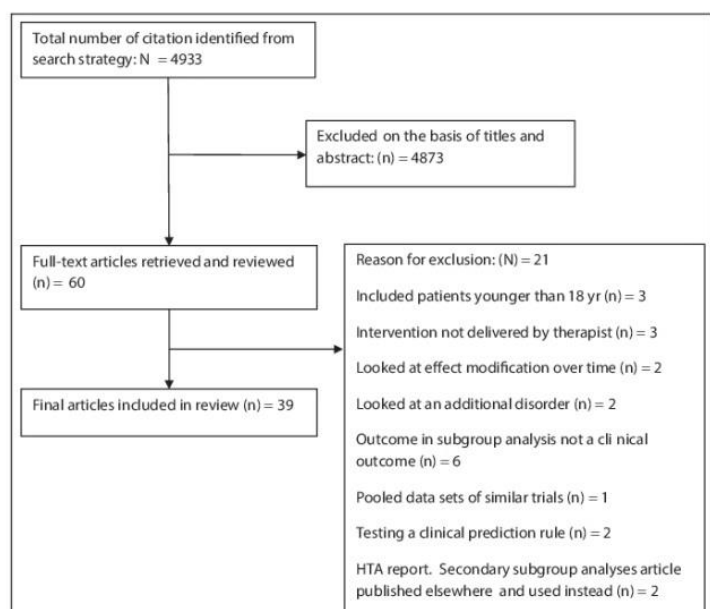


Figure 1. Flow diagram.

TABLE 3. Summary of Included Articles Ordered by Subgroup Quality Assessment							
Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions Compared	Outcome Measure and Follow-up	Subgroups Identified (Only Interaction Test)
Confirmatory findings	Sheets <i>et al</i> ¹⁵	2012	Australia	148	First-line care group vs. McKenzie group	Pain measured at 1 and 3 wk. Global perceived effect at 3 wk.	None
	Smeets <i>et al</i> ¹⁶	2009	Australia and New Zealand	259	Exercise and advice vs. exercise and sham advice vs. sham exercise and sham advice	Pain intensity (11-point scale) and patient-specific function scale score (0–10 scale) measured at baseline, 6 wk, and 52 wk	None
	Underwood <i>et al</i> ¹⁷	2011	UK	701	Advice plus cognitive-behavioral intervention vs. advice alone	RMDQ and MVK scale scores measured at baseline, 3 mo, 6 mo, and 12 mo	Age and employment
Exploratory Findings	Becker <i>et al</i> ¹⁸	2008	Germany	1378	Multifaceted guideline implementation vs. guideline implementation plus motivational counseling vs. postal dissemination of guideline (control)	Hannover Functional Ability Questionnaire administered at baseline and 6 mo	None
	Cecchi <i>et al</i> ¹⁹	2012	Italy	210	Back school vs. individual physiotherapy vs. spinal manipulation	RMDQ score measured at baseline, 3 mo, 6 mo, and 12 mo	None
	Cherkin <i>et al</i> ²⁵	1998	USA	321	Physical therapy vs. chiropractic manipulation vs. educational booklet	Bothersomeness of symptoms and RMDQ score measured at baseline, 4 wk, and 12 wk	Mental health
	Cherkin <i>et al</i> ⁴⁶	2001	USA	262	Chinese acupuncture vs. therapeutic massage vs. self-care education	Bothersomeness of symptoms and RMDQ score measured at baseline, 4 wk, 10 wk, and 1 yr	None
	Cherkin <i>et al</i> ⁴⁷	2009	USA	638	Individualized acupuncture vs. standardized acupuncture vs. simulated acupuncture vs. usual care	Bothersomeness of symptoms and RMDQ score measured at baseline, 8 wk, 26 wk, and 1 yr	None
	Hansen <i>et al</i> ²⁰	1993	Denmark	180	Intensive dynamic back muscle exercise vs. conventional physiotherapy vs. placebo control (semi-hot packs and light traction)	Pain level (10-point scale) measured at baseline, 4 wk, 6 wk, and 1 yr	None
	Hay <i>et al</i> ²¹	2005	UK	402	Brief pain management vs. manual physiotherapy	RMDQ score measured at baseline, 3 mo, and 12 mo	None
	Juni <i>et al</i> ²⁰	2009	Switzerland	104	Standard care alone vs. standard care plus SMT	Pain intensity (11-point scale) and analgesic use measured at baseline, days 1–14, and 6 mo	None

(Continued)

TABLE 3. (Continued)							
Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions Compared	Outcome Measure and Follow-up	Subgroups Identified (Only Interaction Test)
	Karjalainen et al ⁴⁹	2004	Finland	170	Mini intervention group vs. worksite visit group vs. usual care group	Pain intensity (11-point scale) measured at baseline, 3 mo, 6 mo, 1 yr, and 2 yr	Perceived risk for not recovering and type of occupation (comparing mini intervention vs. usual care and worksite visit vs. usual care)
	Kole-Snijders et al ⁵⁰	1999	The Netherlands	159	Operant behavioral treatment with cognitive coping skills training vs. operant behavioral treatment with group discussion vs. waiting list control	Main outcome unclear. Outcomes measured at post-treatment, 6 mo, and 1 yr	None
	Roche et al ⁵¹	2007	France	132	Active individual therapy vs. FRP	Main outcome unclear. Outcomes measured at baseline and 5 wk	Sorensen score
	Sherman et al ⁴³	2009	USA	638	Individualized acupuncture vs. standardized acupuncture vs. simulated acupuncture vs. usual care	Bothersomeness of symptoms and RMDQ score measured at baseline, 8 wk, 26 wk, and 1 yr	Baseline RMDQ
	Smeets et al ⁴⁰	2006	The Netherlands	223	ATP vs. CBT vs. combined ATP and CBT (or combination treatment) vs. WL	RMDQ score measured at baseline, 10 wk, 6 mo, and 12 mo	Baseline RMDQ
	Smeets et al ⁵²	2008	The Netherlands	223	ATP vs. graded activity with problem-solving training vs. combination treatment vs. WL	RMDQ score measured at baseline, 10 wk, 6 mo, and 12 mo	None
	Tilbrook et al ⁶	2011	UK	313	Yoga vs. usual care	RMDQ score measured at baseline, 3 mo, 6 mo, and 12 mo	None
	Underwood et al ⁵⁷	2007	UK	1334	Control (best care in general practice) vs. exercise program vs. spinal manipulation vs. combined treatment (manipulation and exercise)	RMDQ score measured at baseline, 3 mo, and 1 yr	Expectation
	van der Hulst et al ⁴⁴	2008	The Netherlands	163	RRP vs. usual care	RMDQ score measured at baseline, 1 wk after treatment, and 4 mo after treatment	Pain intensity and depression
	Witt et al ¹⁸	2006	Germany	3093	Acupuncture vs. control (delayed acupuncture treatment 3 mo later)	Hannover Functional Ability Questionnaire (0–100 scale) administered at baseline, 3 mo, and 6 mo	Initial back pain, age, and years of schooling
Insufficient Findings	Bendix et al ⁵⁰	1998	Denmark	816	FRP vs. outpatients program (control)	Main outcome unclear. Outcomes measured at baseline and 1 yr	
	Beurskens et al ⁵⁴	1995	The Netherlands	151	Traction vs. sham traction	GPE and severity measured on the VAS at baseline and 5 wk	

(Continued)

TABLE 3. (Continued)

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions Compared	Outcome Measure and Follow-up	Subgroups Identified (Only Interaction Test)
	Bishop et al ²³	2011	USA	112	Supine thrust technique vs. side-lying thrust vs. nonthrust technique	ODQ administered at 1 wk, 4 wk, and 6 mo	None
	Carr et al ²²	2005	UK	237	Group exercise program vs. individual physiotherapy	RMDQ score measured at baseline, 3 mo, and 6 mo	
	Ferreira et al ⁴²	2009	Australia	191	General exercise vs. motor control exercise vs. SMT	GPE (11-point scale), patient-specific functional status, RMDQ score, pain intensity (10-point scale), and spinal stiffness measured at baseline and 8 wk	None
	Glazov ⁴¹	2010	Australia	100	Laser acupuncture vs. sham acupuncture (control)	Pain (VAS) measured at baseline, immediately after treatment, 6 wk, and 6 mo	
	Gudavalli et al ¹⁹	2006	USA	235	Flexion distraction vs. active trunk exercise protocol	Perceived pain (VAS), RMDQ score, and SF-36 score measured at baseline, 4 wk, 3 mo, 6 mo, and 1 yr	
	Hsieh et al ¹⁶	2004	China	146	Acupressure vs. physical therapy	Short-form pain questionnaire administered at baseline, 4 wk, and 6 mo	
	Jellema et al ¹³	2005	The Netherlands	314	Minimal intervention strategy vs. usual care	RMDQ score, perceived recovery (7-point scale), and sick leave measured at baseline, 6 wk, 13 wk, 26 wk, and 1 yr	
	Johnson et al ²⁷	2007	UK	234	Group exercise and education using a cognitive-behavioral approach vs. usual care	Pain (VAS) and RMDQ score measured at baseline, 3 mo, 9 mo, and 15 mo	Patient preference
	Kalauokalani et al ¹⁰	2001	USA	166	Acupuncture vs. massage (subanalysis of the Cherkin et al ¹⁰ article)	RMDQ score measured at baseline, 4 wk, 10 wk, and 1 yr	Patient expectations
	Mellin et al ¹³	1989	Finland	456	Inpatient treatment vs. outpatient treatment vs. control (advice)	Low back pain disability index (scale 0-45) administered at baseline and 3 mo	
	Klaber Moffett et al ¹⁴	2004	UK	187	Exercise vs. usual care	RMDQ score measured at baseline, 6 wk, 6 mo, and 1 yr	
	Myers et al ¹³	2008	USA	444	Usual care vs. usual care plus patient choice of acupuncture, chiropractic, or massage	RMDQ score measured at baseline, 5 wk, and 12 wk	None

(Continued)

TABLE 3. (Continued)

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions Compared	Outcome Measure and Follow-up	Subgroups Identified (Only Interaction Test)
	Seferlis et al ¹⁹	1998	Sweden	180	Manual therapy program vs. intensive training program vs. general practitioner program	Main outcome unclear. Outcomes measured at baseline, 1 mo, 3 mo, and 12 mo	
	Thomas et al ⁴¹	2006	UK	241	Traditional acupuncture vs. usual care	Bodily pain dimension of the SF-36 (0–100 scale) measured at baseline, 3 mo, 12 mo, and 24 mo	Expectation
	van der Roer et al ¹⁵	2008	The Netherlands	114	Intensive group training protocol vs. guideline group	RMDQ score measured at baseline, 6 wk, 13 wk, 26 wk, and 52 wk	
	Vollenbroek-Hutten et al ¹⁶	2004	The Netherlands	163	RRP vs. usual care	RMDQ score measured at baseline, 1 wk after treatment, and 4 mo after treatment	

RMDQ indicates Roland and Morris Disability Questionnaire; MVK, Modified Von Korf (pain and disability); GPE, global perceived effect; ODI, Oswestry Disability Questionnaire; SMT, spinal manipulative therapy; APT, active physical treatment; CBT, cognitive-behavioral treatment; WL, waiting list; RRP, Roessingh back rehabilitation; FRP, functional restoration program; VAS, visual analogue scale; SF-36, 36-item Short Form Health Survey.

catastrophizing, coping, pain self-efficacy, anxiety, depression, stress, troublesomeness, fear avoidance, patient expectation, pain changes with position or movement, presence of leg pain, pain worse with flexion, and duration.^{15–17,19–23} A number of articles did not mention in the “Methods” section that subgroup analyses would be performed and just included the analyses in the “Results” section of the article.^{24–27} All 39 articles, regardless of whether they performed a formal test for interaction, did measure the subgroups of interest before randomization. The majority of these subgroup factors ($n = 33$; 85%) were measured using adequate (reliable and valid) measurements. The commonest subgroup factor that was not measured adequately was patient expectation.

Only one article (3%) gave any indication as to the size and direction of the subgroup effect the authors were expecting before performing the analyses.¹⁵ Three of the articles (8%) provided only a prediction for the direction of the subgroup effect, and around a third of the articles provided some justification regarding the choice of subgroups to be analyzed.

Around a third of articles were able to provide exact definitions of subgroups, although only 5 (13%) provided clear justification for the cutoff points used to define subgroups. All 39 articles performed subgroup analyses on the primary outcome, of which 4 (10%) also performed subgroup analyses using secondary outcomes.

Two articles in particular conducted around 60 interaction tests in addition to the primary analyses, implying a substantial inflation of the overall type I error rate, thus increasing the chance of detecting spurious findings.^{16,28} Of the 3 articles that provided confirmatory findings, only one of them acknowledged and dealt with the issue of multiple testing. They did this by including a multiplicity correction (Bonferroni correction) for the confirmatory subgroup analyses performed.¹⁷

Twelve of the articles (31%) did not use a statistical test for interaction to assess for treatment effect modification. Two of these articles did not give any indication as to what statistical method they used for this.^{24,29} Two articles looked at correlations between individual subgroups and outcomes within each treatment arm separately.^{30,31} Two articles used t tests between treatment groups within individual subgroups.^{19,26} Five articles used either multiple linear regression or multiple logistic regression for each individual subgroup.^{32–36} Finally, 1 article compared the medians across 3 trial arms within individual subgroups using Kruskal-Wallis tests.²⁰

Assessment of Conduct and Reporting: Reporting of Results

There is some confusion in the articles between investigating “subgroup effects” and investigating “differential subgroup effects.”³⁷ Twenty-seven articles (69%) used a statistical test for interaction to perform subgroup analyses. Four of these reported subgroup analyses within individual subgroups,^{38–41} 10 articles reported results from interaction tests,^{15–17,22,23,28,42–45} and the remaining 13 articles either did not report any results at all or just reported the P value for the interaction term.^{5,18,20,21,25,27,46–52} Six articles reported both the interaction effect sizes with confidence intervals and the corresponding

TABLE 4. Summary of Excluded Articles	
Article	Reason for Exclusion
Childs JD, Flynn TW, Fritz JM. A perspective for considering the risks and benefits of spinal manipulation in patients with low back pain. <i>Man Ther</i> 2006;11:316–20	Testing a clinical prediction rule
Costa LO, Maher CG, Latimer J, et al. Motor control exercise for chronic low back pain: a randomized placebo-controlled trial. <i>Phys Ther</i> 2009;89:1275–86	Look at effect modification over time
Faas A, Chavannes AW, van Eijk JT, et al. A randomized, placebo-controlled trial of exercise therapy in patients with acute low back pain. <i>Spine</i> 1993;18:1388–95	Included patients younger than 18 yr
Faas A, van Eijk JT, Chavannes AW, et al. A randomized trial of exercise therapy in patients with acute low back pain. Efficacy on sickness absence. <i>Spine</i> 1995;20:941–7	Included patients younger than 18 yr and outcome in subgroup analyses not a clinical measure of low back pain (sickness absence)
George SZ, Fritz JM, Childs JD, et al. Sex differences in predictors of outcome in selected physical therapy interventions for acute low back pain. <i>J Orthop Sports Phys Ther</i> 2006;36:354–63	Pooled data sets of similar trials
George SZ, Zeppieri G Jr, Cere AL, et al. A randomized trial of behavioral physical therapy interventions for acute and sub-acute low back pain (NCT00373867). <i>Pain</i> 2008;140:145–57	Included patients younger than 18 yr and also looked at effect modification over time
Haas M, Grouppe E, Muench J, et al. Chronic disease self-management program for low back pain in the elderly. <i>J Manip Physiol Ther</i> 2005;28:228–37	Intervention not delivered by therapist
Hagen EM, Svensen E, Eriksen HR. Predictors and modifiers of treatment effect influencing sick leave in subacute low back pain patients. <i>Spine</i> 2005;30:2717–23	Outcome in subgroup analyses not a clinical measure of low back pain (return to work)
Hancock MJ, Maher CG, Latimer J, et al. Independent evaluation of a clinical prediction rule for spinal manipulative therapy: a randomised controlled trial. <i>Eur Spine J</i> 2008;17:936–43	Testing a clinical prediction rule
Jellema P, van der Windt DA, van der Horst HE, et al. Should treatment of (sub) acute low back pain be aimed at psychosocial prognostic factors? Cluster randomised clinical trial in general practice. <i>BMJ</i> 2005;331:84	Look at effect modification over time
Jellema P, van der Roer N, van der Windt DA, et al. Low back pain in general practice: cost-effectiveness of a minimal psychosocial intervention versus usual care. <i>Eur Spine J</i> 2007;16:1812–21	Outcome in subgroup analyses not a clinical measure of low back pain (cost-effectiveness)
Kool JP, Oesch PR, Bachmann S, et al. Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. <i>Arch Phys Med Rehabil</i> 2005;86:857–64	Outcome in subgroup analyses not a clinical measure of low back pain (days worked >3 mo)
Lamb SE, Lall R, Hansen Z, et al. A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. <i>Health Technol Assess (Winchester, England)</i> 2000;14:1–253	HTA report. Secondary subgroup analysis article published elsewhere and used instead (Underwood et al ¹¹)
Scheel IB, Hagen KB, Herrin J, et al. A randomized controlled trial of two strategies to implement active sick leave for patients with low back pain. <i>Spine</i> 2002;27:561–6	Outcome in subgroup analyses not a clinical measure of low back pain (active sick leave)
Skargren EI, Carlsson PG, Oberg BE. One-year follow-up comparison of the cost and effectiveness of chiropractic and physiotherapy as primary management for back pain. Subgroup analysis, recurrence, and additional health care utilization. <i>Spine</i> 1998;23:1875–83	Looked at an addition disorder (neck pain)
Skargren EI, Oberg BE, Carlsson PG, et al. Cost and effectiveness analysis of chiropractic and physiotherapy treatment for low back and neck pain. Six-month follow-up. <i>Spine</i> 1997;22:2167–77	Looked at an addition disorder (neck pain)
Staal JB, Hlobil H, Koke AJ, et al. Graded activity for workers with low back pain: who benefits most and how does it work? <i>Arthritis Rheum</i> 2008;59:642–9	Outcome in subgroup analyses not a clinical measure of low back pain (return to work)
Steenstra IA, Knol DL, Bongers PM, et al. What works best for whom? An exploratory, subgroup analysis in a randomized, controlled trial on the effectiveness of a workplace intervention in low back pain patients on return to work. <i>Spine</i> 2009;34:1243–9	Outcome in subgroup analyses not a clinical measure of low back pain (return to work)

(Continued)

TABLE 4. (Continued)

Article	Reason for Exclusion
Thomas KJ, MacPherson H, Ratcliffe J, et al. Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain. <i>Health Technol Assess (Winchester, England)</i> 2010;9:iii–iv	HTA report. Secondary subgroup analysis article published elsewhere and used instead (Thomas et al ⁴¹)
Toda Y. Impact of waist/hip ratio on the therapeutic efficacy of lumbosacral corsets for chronic muscular low back pain. <i>J Orthop Sci</i> 2002;7:644–9	Intervention not delivered by therapist (corsets given to patients)
van Poppel MN, Koes BW, van der Ploeg T, et al. Lumbar supports and education for the prevention of low back pain in industry: a randomized controlled trial. <i>JAMA</i> 1998;279:1789–94	Intervention not delivered by therapist (lumbar supports given to patients)

HTA indicates health technology assessment.

P values,^{23,28,42–45} 4 articles reported only the interaction effect sizes with confidence intervals,^{15–17,22} 8 articles reported only the *P* values,^{5,18,27,39,41,48,49,51} and 9 articles did not report either the interaction effect sizes and confidence intervals or *P* values.^{20,21,25,38,40,46,47,50,52}

Assessment of Conduct and Reporting: Interpretation and Discussion

Four of 27 articles that performed interaction tests reported subgroup analyses within individual subgroups and thus based the interpretations and discussion on this as well. Around a third of the articles provided supporting or contradictory findings from other relevant studies. Twelve of the 27 articles that used an interaction test reported significant findings, of which only 2 suggested the identified subgroups for investigation in future studies. Twelve articles acknowledged the limitations of performing subgroup analyses in the “Discussion” section of the article.

DISCUSSION

Our aim in this review was to assess the quality, conduct, and reporting of subgroup analyses performed in RCTs of therapist-delivered interventions for the management of NSLBP. We think this is the first study of the overall quality, conduct, and reporting of subgroup analyses in the area of low back pain.

Reporting Quality

Many authors have performed subgroup analyses or have attempted some form of subgroup analyses. There is some confusion between investigating “subgroup effects” and investigating “differential subgroup effects.”³⁷ The results of the quality assessment suggest that only 3 of the articles are able to provide confirmatory findings, and that the majority of the articles provide exploratory or insufficient findings. These results are solely based on the outcomes from the quality assessment; the results of an apparently high-quality subgroup analysis need to be interpreted in light of the quality of the main study. The general content and reporting of these articles in relation to subgroup analyses, that is, in terms of design and methods, results, interpretation, and discussion, are rather quite poor. These articles can be seen as missed opportunities. Several of these subgroup analysis articles

could have used more appropriate methodology, that is, statistical test for interaction, improved the standard of reporting, or both. Had this been done, they would have been able to contribute valuable information to the existing pool of subgroup-related literature in the area of low back pain. That nearly half of the identified articles provided insufficient findings raises concerns that the already published subgroup literature be misinterpreted when considering future subgroups research or making treatment choice. Although most subgroup analyses lack power and are of an exploratory nature, a well-conducted and reported subgroup analysis will ensure credibility of findings that can be tested in future studies.

Sample Size

All but one of the articles that were reviewed had inadequate sample size and were thus substantially underpowered to detect any meaningful interaction effects in the primary outcome. The one trial designed and powered to detect important treatment-covariate interactions did not, however, prespecify subgroups of interest.¹⁸ Some subgroups were found to significantly moderate treatment effect in this trial; however, it was quite disappointing to see only *P* values reported.

Lack of power is a well-known issue associated with subgroup analyses. Although the power of a particular study depends on the outcome used and the size of any clinically meaningful effect as well as distributional assumptions, based on the simple model for a normally distributed response proposed by Lachenbruch,³³ a total sample size of approximately 500 participants provides 80% power to detect an interaction effect with a moderate standardized effect size of 0.5 at a 2-sided 5% significance level. To date, we are only aware of 4 trials with subgroup analyses that have this sample size ($n \geq 500$).^{6,18,47,54} These were high-quality RCTs that were designed and powered to detect a standardized difference of around 0.4 in the main effect. However, inspection of the main effect sizes suggests that the standardized effects were much lower and only ranged from 0.12 to 0.23.⁵ If the larger trials are failing to pick up a moderate standardized main effect size, it is highly unlikely that any plausible and clinically meaningful interaction effects will be detected in any subgroup analyses unless the interaction effect is considerably larger than the main effect. In a simple model, with equal numbers in each subgroup, such a large interaction effect

would consequentially mean that there is a large benefit in one subgroup and a smaller harm in the other subgroup.

A trial to identify a differential subgroup effect needs to be approximately 4 times larger than a trial powered to detect a main effect of only the same magnitude.⁵⁵ Any such trial would only be able to test a moderator of treatment effect for one subgroup. Unless there is an overwhelming *a priori* hypothesis that needs testing in such a study, this is unlikely to be a worthwhile expenditure of academic effort and funders' resources. We are not aware of any such overwhelming *a priori* hypothesis. Furthermore, our existing pool of baseline predictors explain only about a third of the variance in outcome, making it unlikely that we can identify a single, strong moderator of treatment effect to underpin such a strong *a priori* hypothesis.⁵⁶ Even if such a study was designed, it would, if it was to inform clinical practice, need to ensure that whatever moderators were proposed could easily be applied in clinical practice. Therefore, it is clear that different approaches are needed. We suggest 3 alternative approaches that the research community should consider:

1. Developing phenotypically defined subgroups based on clinical reasoning and developing interventions specifically targeting these groups. One might expect that such targeted interventions would, if effective, have larger average effect sizes than current nonspecific interventions for NSLBP. There would be no justification for deciding whether or not the targeted intervention worked in other phenotypic groups. A series of such trials could be run for the same cost as one mega-trial testing for an interaction.
2. Typical subgroup analyses use interaction tests to investigate patient characteristics one at a time; however, it is quite obvious that patients have multiple characteristics that also need to be investigated either simultaneously or in some stepwise fashion. Novel statistical approaches might allow the identification of multiple patient characteristics or clusters of moderators that would identify who is most (or least) likely to benefit.^{57–59}
3. Where data are available from existing trial data sets that are sufficiently homogeneous, then individual patient data meta-analysis would allow additional insights. This would address the issue of statistical power, but the issue of multiple testing and hence the inflation of the familywise type I error rate (probability of a false-positive result) as well as the issue of little of the variance in outcome explained still remain.

CONCLUSION

Finding moderators of treatment effect has been identified as a high research priority internationally for the management of NSLBP in a step toward better individualized patient care. The findings of this review suggest that the majority of subgroup analyses performed in low back pain trials to date are only able to provide exploratory or insufficient findings. Articles with insufficient findings are not very credible at all and could potentially provide false implications for guiding

future research. Moreover, the general content and reporting of subgroup analyses are rather poor. It is thus recommended that authors use available guidelines when performing subgroup analyses to ensure that they are reliable and of a good standard.^{13,60} We do, however, have serious concerns that current approaches are inadequate for the task at hand. There is a need to find and develop alternative statistical methods for performing subgroup analyses to overcome or better deal with the existing issues associated with current methodology.

Key Points

- ❑ The identification of subgroups that gain the most benefit from interventions for the management of NSLBP is an important research priority internationally.
- ❑ Subgroup analyses performed in NSLBP trials have been severely underpowered, are only able to provide exploratory or insufficient findings, and have poor quality of reporting.
- ❑ There is a need to develop new approaches to subgroup identification in back pain research.

Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (www.spinejournal.com).

References

1. Dagenais S, Caro J, Haldeman S. A systematic review of low back pain cost of illness studies in the United States and internationally. *Spine J* 2008;8:8–20.
2. Maniadakis N, Gray A. The economic burden of back pain in the UK. *Pain* 2000;84:95–103.
3. Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380:2163–96.
4. Savigny P KS, Watson P, Underwood M, et al. *Low Back Pain: Early Management of Persistent Non-Specific Low Back Pain*. London: National Collaborating Centre for Primary Care and Royal College of General Practitioners; 2009.
5. Tilbrook HE, Cox H, Hewitt CE, et al. Yoga for chronic low back pain. *Ann Intern Med* 2011;155:569–78.
6. Team UBT. United Kingdom Back Pain Exercise and Manipulation (UK BEAM) randomised trial: effectiveness of physical treatments for back pain in primary care. *BMJ* 2004;329:1377.
7. Lamb SE, Lall R, Hansen Z, et al. A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. *Health Technol Assess* 2010;14:1–253, iii–iv.
8. Little P, Lewith G, Webley F, et al. Randomised controlled trial of Alexander technique lessons, exercise, and massage (ATEAM) for chronic and recurrent back pain [reprint in: *Br J Sports Med* 2008;42:965–8]. *BMJ* 2008;337:a884.
9. Kraemer HC, Stice E, Kazdin A, et al. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *Am J Psychiatry* 2001;158:848–56.
10. Kent P, Keating J, Leboeuf-Yde C. Research methods for subgrouping low back pain. *BMC Med Res Methodol* 2010;10:62.
11. Borkan JM, Koes B, Reis S, et al. A report from the second international forum for primary care research on low back pain. Reexamining priorities. *Spine (Phila Pa 1976)* 1998;23:1992–6.

12. Pincus T, Miles C, Froud R, et al. Methodological criteria for the assessment of moderators in systematic reviews of randomised controlled trials: a consensus study. *BMC Med Res Methodol* 2011;11:14.
13. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;365:176–86.
14. Lagakos SW. The challenge of subgroup analyses—reporting without distorting. *N Engl J Med* 2006;354:1667–9.
15. Sheets C, Machado LA, Hancock M, et al. Can we predict response to the McKenzie method in patients with acute low back pain? A secondary analysis of a randomized controlled trial. *Eur Spine J* 2012;21:1250–6.
16. Smeets RJEM, Maher CG, Nicholas MK, et al. Do psychological characteristics predict response to exercise and advice for subacute low back pain? *Arthritis Rheum* 2009;61:1202–9.
17. Underwood M, Mistry D, Lall R, et al. Predicting response to a cognitive-behavioral approach to treating low back pain: secondary analysis of the BeST data set. *Arthritis Care Res* 2011;63:1271–9.
18. Witt CM, Jena S, Selim D, et al. Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. *Am J Epidemiol* 2006;164:487–96.
19. Gudavalli MR, Cambron JA, McGregor M, et al. A randomized clinical trial and subgroup analysis to compare flexion-distraction with active exercise for chronic low back pain. *Eur Spine J* 2006;15:1070–82.
20. Hansen FR, Bendix T, Skov P, et al. Intensive, dynamic back-muscle exercises, conventional physiotherapy, or placebo-control treatment of low-back pain. A randomized, observer-blind trial. *Spine (Phila Pa 1976)* 1993;18:98–108.
21. Hay EM, Mullis R, Lewis M, et al. Comparison of physical treatments versus a brief pain-management programme for back pain in primary care: a randomised clinical trial in physiotherapy practice. *Lancet* 2005;365:2024–30.
22. Johnson RE, Jones GT, Wiles NJ, et al. Active exercise, education, and cognitive behavioral therapy for persistent disabling low back pain: a randomized controlled trial. *Spine (Phila Pa 1976)* 2007;32:1578–85.
23. Bishop MD, Bialosky JE, Cleland JA. Patient expectations of benefit from common interventions for low back pain and effects on outcome: secondary analysis of a clinical trial of manual therapy interventions. *J Man Manip Ther* 2011;19:20–5.
24. Beurskens AJ, de Vet HC, Koke AJ, et al. Efficacy of traction for non-specific low back pain: a randomised clinical trial. *Lancet* 1995;346:1596–600.
25. Cherkin DC, Deyo RA, Battie M, et al. A comparison of physical therapy, chiropractic manipulation, and provision of an educational booklet for the treatment of patients with low back pain. *N Engl J Med* 1998;339:1021–9.
26. Hsieh LL-C, Kuo C-H, Yen M-F, et al. A randomized controlled clinical trial for low back pain treated by acupressure and physical therapy. *Prev Med* 2004;39:168–76.
27. Roche G, Ponthieux A, Parot-Shinkel E, et al. Comparison of a functional restoration program with active individual physical therapy for patients with chronic low back pain: a randomized controlled trial. *Arch Phys Med Rehabil* 2007;88:1229–35.
28. Underwood MR, Morton V, Farrin A, et al. Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM dataset [ISRCTN32683578]. *Rheumatology (Oxford)* 2007;46:1297–302.
29. Seferlis T, Nemeth G, Carlsson AM, et al. Conservative treatment in patients sick-listed for acute low-back pain: a prospective randomised study with 12 months' follow-up. *Eur Spine J* 1998;7:461–70.
30. Bendix AF, Bendix T, Hastrup C. Can it be predicted which patients with chronic low back pain should be offered tertiary rehabilitation in a functional restoration program? A search for demographic, socioeconomic, and physical predictors. *Spine (Phila Pa 1976)* 1998;23:1775–83; discussion 83–4.
31. Mellin G, Hurri H, Harkapaa K, et al. A controlled study on the outcome of inpatient and outpatient treatment of low back pain, part II: effects on physical measurements three months after treatment. *Scand J Rehabil Med* 1989;21:91–5.
32. Carr JL, Klaber Moffett JA, Howarth E, et al. A randomized trial comparing a group exercise programme for back pain patients with individual physiotherapy in a severely deprived area. *Disabil Rehabil* 2005;27:929–37.
33. Jellema P, van der Windt DAWM, van der Horst HE, et al. Why is a treatment aimed at psychosocial factors not effective in patients with (sub)acute low back pain? *Pain* 2005;118:350–9.
34. Klaber Moffett JA, Carr JM, Howarth EM. High fear-avoiders of physical activity benefit from an exercise program for patients with back pain. *Spine (Phila Pa 1976)* 2004;29:1167–72.
35. van der Roer N, van Tulder M, Barendse J, et al. Intensive group training protocol versus guideline physiotherapy for patients with chronic low back pain: a randomised controlled trial. *Eur Spine J* 2008;17:193–200.
36. Vollenbroek-Hutten MMR, Hermens HJ, Wever D, et al. Differences in outcome of a multidisciplinary treatment between subgroups of chronic low back pain patients defined using two multi-axial assessment instruments: the multidimensional pain inventory and lumbar dynamometry. *Clin Rehabil* 2004;18:566–79.
37. Yusuf S, Wittes J, Probstfield J, et al. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266:93–8.
38. Becker A, Leonhardt C, Kochen MM, et al. Effects of two guideline implementation strategies on patient outcomes in primary care: a cluster randomized controlled trial. *Spine (Phila Pa 1976)* 2008;33:473–80.
39. Juni P, Battaglia M, Nuesch E, et al. A randomised controlled trial of spinal manipulative therapy in acute low back pain. *Ann Rheum Dis* 2009;68:1420–7.
40. Smeets RJEM, Vlaeyen JWS, Hidding A, et al. Active rehabilitation for chronic low back pain: cognitive-behavioral, physical, or both? First direct post-treatment results from a randomized controlled trial [ISRCTN22714229]. *BMC Musculoskelet Disord* 2006;7:5.
41. Thomas KJ, MacPherson H, Thorpe L, et al. Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. *BMJ* 2006;333:623.
42. Ferreira ML, Ferreira PH, Latimer J, et al. Relationship between spinal stiffness and outcome in patients with chronic low back pain. *Man Ther* 2009;14:61–7.
43. Sherman KJ, Cherkin DC, Ichikawa L, et al. Characteristics of patients with chronic back pain who benefit from acupuncture. *BMC Musculoskelet Disord* 2009;10:114.
44. van der Hulst M, Vollenbroek-Hutten MMR, Groothuis-Oudshoorn KGM, et al. Multidisciplinary rehabilitation treatment of patients with chronic low back pain: a prognostic model for its outcome. *Clin J Pain* 2008;24:421–30.
45. Cecchi F, Negrini S, Pasquini G, et al. Predictors of functional outcome in patients with chronic low back pain undergoing back school, individual physiotherapy or spinal manipulation. *Eur J Phys Rehabil Med* 2012;48:371–8.
46. Cherkin DC, Eisenberg D, Sherman KJ, et al. Randomized trial comparing traditional Chinese medical acupuncture, therapeutic massage, and self-care education for chronic low back pain. *Arch Intern Med* 2001;161:1081–8.
47. Cherkin DC, Sherman KJ, Avins AL, et al. A randomized trial comparing acupuncture, simulated acupuncture, and usual care for chronic low back pain. *Arch Intern Med* 2009;169:858–66.
48. Kalauokalani D, Cherkin DC, Sherman KJ, et al. Lessons from a trial of acupuncture and massage for low back pain: patient expectations and treatment effects. *Spine (Phila Pa 1976)* 2001;26:1418–24.
49. Karjalainen K, Malmivaara A, Mutanen P, et al. Mini-intervention for subacute low back pain: two-year follow-up and modifiers of effectiveness. *Spine (Phila Pa 1976)* 2004;29:1069–76.
50. Koe-Snijders AM, Vlaeyen JW, Goossens ME, et al. Chronic low-back pain: what does cognitive coping skills training add to operant

- behavioral treatment? Results of a randomized clinical trial. *J Consult Clin Psychol* 1999;67:931–44.
51. Myers SS, Phillips RS, Davis RB, et al. Patient expectations as predictors of outcome in patients with acute low back pain. *J Gen Intern Med* 2008;23:148–53.
52. Smeets RJE, Vlaeyen JWS, Hidding A, et al. Chronic low back pain: physical training, graded activity with problem solving training, or both? The one-year post-treatment results of a randomized controlled trial [reprint in: *Ned Tijdschr Geneesk* 2009;153:543–9]. *Pain* 2008;134:263–76.
53. Lachenbruch PA. A note on sample size computation for testing interactions. *Stat Med* 1988;7:467–9.
54. Lamb SE, Hansen Z, Lall R, et al. Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis. *Lancet* 2010;375:916–23.
55. Brookes ST, Whitley E, Peters TJ, et al. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess* 2001;5:1–56.
56. Pincus T, Santos R, Breen A, et al. A review and proposal for a core set of factors for prospective cohorts in low back pain: a consensus statement. *Arthritis Care Res* 2008;59:14–24.
57. Su X, Tsai C-L, Wang H, et al. Subgroup analysis via recursive partitioning. *SSRN eLibrary* 2009;10:141–58.
58. Dusseldorp E, Meulman JJ. The regression trunk approach to discover treatment covariate interaction. *Psychometrika* 2004;69:355–74.
59. Lipkovich I, Dmitrienko A, Denne J, et al. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med* 2011;30:2601–21.
60. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007;357:2189–94.
61. Glazov G. The influence of baseline characteristics on response to a laser acupuncture intervention: an exploratory analysis. *Acupunct Med* 2010;28:6–11.

Appendix B

Appendix B: Systematic review search terms

Listed below are the specific keywords used to conduct the literature search:

- 1 Low Back Pain/ (11231)
- 2 "low* back pain".m_titl. (6615)
- 3 1 or 2 (12932)
- 4 randomized controlled trial.pt. (309096)
- 5 controlled clinical trial.pt. (82630)
- 6 randomized.ab. (215472)
- 7 clinical trials as topic.sh. (154789)
- 8 randomly.ab. (155983)
- 9 trial.ti. (92244)
- 10 4 or 5 or 6 or 7 or 8 or 9 (693954)
- 11 3 and 10 (1993)

Appendix C

Appendix C: Matrix form of the general expressions required to generate data for the simulation study

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{pmatrix} = \begin{pmatrix} pq_1r & pq_1(1-r) & p(1-q_1)r & p(1-q_1)(1-r) & (1-p)q_2r & (1-p)q_2(1-r) & (1-p)(1-q_2)r & (1-p)(1-q_2)(1-r) \\ pq_1 & -pq_1 & p(1-q_1) & -p(1-q_1) & (1-p)q_2 & -(1-p)q_2 & (1-p)(1-q_2) & -(1-p)(1-q_2) \\ rq_1 & (1-r)q_1 & r(1-q_1) & (1-r)(1-q_1) & -rq_2 & -(1-r)q_2 & -r(1-q_2) & -(1-r)(1-q_2) \\ pr & p(1-r) & -pr & -p(1-r) & (1-p)r & (1-p)(1-r) & -(1-p)r & -(1-p)(1-r) \\ r & (1-r) & -r & -(1-r) & -r & -(1-r) & r & (1-r) \\ q_1 & -q_1 & (1-q_1) & -(1-q_1) & -q_2 & q_2 & -(1-q_2) & (1-q_2) \\ p & -p & -p & p & (1-p) & -(1-p) & -(1-p) & (1-p) \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix} \times \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{pmatrix}$$